


3 1761 10374383 7



Digitized by the Internet Archive
in 2023 with funding from
University of Toronto

<https://archive.org/details/31761103743837>

12-001



Government
Publications

224

SURVEY METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

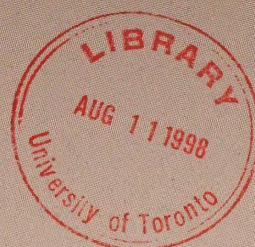
JUNE 1998

•

VOLUME 24

•

NUMBER 1



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL PUBLISHED BY STATISTICS CANADA

JUNE 1998 • VOLUME 24 • NUMBER 1

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 1998

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

July 1998

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics Canada
Statistique Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman	G.J. Brackstone	
Members	D. Binder	R. Platek (Past Chairman)
	G.J.C. Hole	D. Roy
	F. Mayda (Production Manager)	M.P. Singh
	C. Patrick	

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, <i>University of Western Ontario</i>	D. Pfeffermann, <i>Hebrew University</i>
D. Binder, <i>Statistics Canada</i>	J.N.K. Rao, <i>Carleton University</i>
J.-C. Deville, <i>INSEE</i>	L.-P. Rivest, <i>Université Laval</i>
J.D. Drew, <i>Statistics Canada</i>	I. Sande, <i>Bell Communications Research, U.S.A.</i>
J. Eltinge, <i>Texas A&M University</i>	F.J. Scheuren, <i>Ernst and Young, LLP</i>
W.A. Fuller, <i>Iowa State University</i>	J. Sedransk, <i>Case Western Reserve University</i>
R.M. Groves, <i>University of Maryland</i>	R. Sitter, <i>Simon Fraser University</i>
M.A. Hidioglou, <i>Statistics Canada</i>	C.J. Skinner, <i>University of Southampton</i>
D. Holt, <i>Central Statistical Office, U.K.</i>	R. Valliant, <i>Westat, Inc.</i>
G. Kalton, <i>Westat, Inc.</i>	V.K. Verma, <i>University of Essex</i>
R. Lachapelle, <i>Statistics Canada</i>	P.J. Waite, <i>U.S. Bureau of the Census</i>
P. Lahiri, <i>University of Nebraska-Lincoln</i>	J. Waksberg, <i>Westat, Inc.</i>
S. Linacre, <i>Australian Bureau of Statistics</i>	K.M. Wolter, <i>National Opinion Research Center</i>
G. Nathan, <i>Central Bureau of Statistics, Israel</i>	A. Zaslavsky, <i>Harvard University</i>

Assistant Editors J. Denis, P. Dick, H. Mantel and D. Stukel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is \$47 per year in Canada and US \$47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 24, Number 1, June 1998

CONTENTS

In This Issue	1
P.S. KOTT, J.F. AMRHEIN and S.D. HICKS Sampling and Estimation From Multiple List Frames	3
M.A. HIDIROGLOU and C.-E. SÄRNDAL Use of Auxiliary Information for Two-phase Sampling	11
T.L. BYCZKOWSKI, M.S. LEVY and D.J. SWEENEY Estimation in Sample Surveys Using Frames With a Many-to-Many Structure	21
I.S. YANSANEH and W.A. FULLER Optimal Recursive Estimation for Repeated Surveys	31
S. SINGH, S. HORN and F. YU Estimation of Variance of General Regression Estimator: Higher Level Calibration Approach	41
R. LEHTONEN and A. VEIJANEN Logistic Generalized Regression Estimators	51
R.J. CASADY, A.H. DORFMAN and S. WANG Confidence Intervals for Domain Parameters When the Domain Sample Size is Random	57
G.E. MONTANARI On Regression Estimation of Finite Population Means	69
D.E. HAINES and K.H. POLLOCK Combining Multiple Frames to Estimate Population Size and Totals	79
N. BATES and E.R. GERBER Temporary Mobility and Reporting of Usual Residence	89

In This Issue

This issue of *Survey Methodology* contains articles on a variety of topics. Kott, Amrhein and Hicks tackle the problem of multi-purpose surveys. For such surveys, it would be desirable to be able to stratify the target population in various ways in order to improve the precision of the estimates of interest. The authors present four sampling methods for the selection of samples through various stratifications while reducing the overall size of the sample. These strategies are then evaluated using data taken from an agriculture survey. The authors then show how a calibration estimator can improve the relative efficiency.

Singh, Horn and Yu examine the problem of estimating the variance of the general linear regression estimator. They carry out calibration at two distinct levels. The higher-level calibration thus defined uses the known total and variance of the auxiliary variables. The authors show that this method covers a broader range of estimators than the lower-level calibration method, which uses only the known total of the auxiliary variables. An empirical study is presented to assess the efficiency of the proposed strategies.

Hidirolou and Särndal concern themselves with the use of auxiliary data in two-phase sampling. They explain how these data are converted into calibration weight, in two phases, in order to create efficient estimators of a population total. The authors show that the calibration estimator, using the generalized least squares function, can be expressed as a perfectly equivalent two-phase regression estimator, that is, an estimator that is the product of two successive regression fits. They examine forms of the two-phase calibration estimator when the auxiliary data are for population subsets known as "calibration groups." They also discuss the estimation of domains of interest and the estimation of variance.

Byczkowski, Levy and Sweeney consider survey frames having a many-to-many structure, that is, any unit in the frame may be associated with multiple target population elements and any target population element may be associated with multiple frame units. This problem is motivated by a building characteristics survey in which the target population consists of commercial buildings, but the frame consists of a list of street addresses (which in turn correspond to either single buildings, multiple buildings or parts of buildings). Under this setting, estimators of totals and means and their variances using simple and stratified random sampling without replacement are developed.

Yansaneh and Fuller present a recursive regression estimation procedure to reduce the computational complexity associated with best linear unbiased estimation in the context of a repeated survey with partial overlap. They use data from the U.S. Current Population Survey (CPS) to compare variances of their recursive regression estimator to some alternative estimators including the current CPS composite estimator. The proposed estimator seems to be very competitive for estimates of both level and change. They also estimate variances under various rotation patterns and find that the current 4-8-4 rotation pattern is superior to continuous rotation for current level and long-period averages, but inferior for short period changes.

Lehtonen and Veijanen bring together two well-known ideas, generalized regression (GREG) and pseudo maximum likelihood estimation, to develop a new methodology for estimating the population total of a categorical survey variable, given a vector of known auxiliary variables. The values of the categorical variable are modeled as realizations from a multinomial logistic and the corresponding unknown parameters are estimated through pseudo maximum likelihood. Then, the population frequencies of interest are estimated via a modified GREG estimator which uses these estimated parameters. Variance estimates of the frequencies are given through Taylor linearization, and some empirical results based on Finnish Labour Force Survey data are provided.

Casady, Dorfman and Wang consider the construction of confidence intervals for domain parameters in the case where the domain sample size is not fixed by the design. They condition on the observed domain sample size and show how, under certain assumptions about the population, conditional t -based confidence intervals can be obtained. In an empirical study using data from the U.S. Bureau of Labor Statistics Occupational Compensation survey, they demonstrate that the proposed conditional intervals have better coverage probabilities than standard marginal intervals.

Montanari compares two well-known estimators of a finite population mean: the GREG and the design-optimal regression estimator obtained from the difference estimator. While the former can be inefficient if the underlying model is misspecified, the latter, although model-free, is vulnerable to sampling fluctuations. An efficiency measure, which provides a criterion for choosing between the two estimators, is given. The results of an empirical study, which investigates the behaviour of both estimators under a variety of misspecified and correct models, are discussed.

Haines and Pollock provide a fresh examination of estimating totals with multiple frames. Estimators are developed when information is only available from list frames and, in addition, when information is also provided from an area frame. A simulation shows that the best estimator depends on the known, or assumed, dependence of the frames. They also study the situation when observations are either available for all units or only available for a sub-sample from each frame. Again, the preferred estimator changes when the dependence between frames is considered.

Bates and Gerber analyze the dynamics of a difficult problem: how temporary mobility of an individual contributes to within-household coverage error. They develop a two dimensional typology to characterize temporary mobility, then using data from the Living Situation Survey, conducted in the U.S. in 1993, they identify four temporary mobility patterns. Two of these traits are found to be useful predictors of persons missed from censuses or surveys.

The Editor

Sampling and Estimation From Multiple List Frames

PHILLIP S. KOTT, JOHN F. AMRHEIN and SUSAN D. HICKS¹

ABSTRACT

Many economic and agricultural surveys are multi-purpose. It would be convenient if one could stratify the target population of such a survey in a number of different ways to satisfy a number of different purposes and then combine the samples for enumeration. We explore four different sampling methods that select similar samples across all stratifications thereby reducing the overall sample size. Data from an agriculture survey is used to evaluate the effectiveness of these alternative sampling strategies. We then show how a calibration (*i.e.*, reweighted) estimator can increase statistical efficiency by capturing what is known about the original stratum sizes in the estimation. Raking, which has been suggested in the literature for this purpose, is simply one method of calibration.

KEY WORDS: Calibration; Collocated sampling; Permanent random numbers; Poisson sampling; Systematic probability proportional to size sampling.

1. INTRODUCTION

Many of the list frame surveys conducted by the National Agricultural Statistics Service (NASS) are integrated in the sense that data on a range of heterogeneous items, such as planted crop acres and grain stock inventories, are collected in a single survey rather than through a number of independent surveys. Bankier (1986), Skinner (1991), and Skinner, Holmes and Holt (1994) have shown how an old method of combining independently drawn stratified simple random samples – where each sample comes from a (list) frame with a different stratification scheme – can be made more efficient; that is, the variances resulting from such a combined estimation strategy would not be as large as those from the independent surveys summarized by themselves.

Even more appealing for many applications would be a sampling design that tends to select the same units from every frame, thereby reducing both the cost and respondent burden of an integrated survey. This paper explores several such designs. Three make use of permanent random numbers. The fourth uses a variation of systematic probability proportional to size sampling. The goal for each is to meet or exceed – at least on average – a particular set of sample size targets.

The paper shows how a calibration (*i.e.*, reweighted) estimator can provide relative efficiency by capturing what we know about the original stratum sizes in the estimation. A final section points out that the use of a calibration technique can do more than simply reflect original stratum sizes.

An alternative strategy for burden reduction is to use separate instruments for different survey targets and to select distinct samples for each instrument. This increases the number of units selected over all, but reduces the burden per selected unit. NASS is using that approach in its Agricultural Resources Management Study (see Kott and Fetter 1997), but it is *not* the approach to be discussed here.

2. INDEPENDENT SAMPLING AND UNBIASED ESTIMATION

Suppose we have F independent frames; for example, a sorghum frame, an oats frame, and a general grain stocks frame. Each frame is stratified independently, and without replacement simple random samples are drawn from each stratum of every frame. Frame f (say, the oats frame) contains H_f strata; stratum h (large oats operations) in frame f has N_{fh} population units, out of which n_{fh} units are selected. The union of the F frames must cover the entire (list) population, but no single frame need be complete. The frames may overlap.

One unbiased estimator for a population total $T = \sum_{i \in P} y_i$ is the simple multiplicity estimator suggested by Skinner (1991):

$$t_M = \sum_{i \in P} y_i n_{(i)} / E[n_{(i)}], \quad (1)$$

here P denotes the entire population, and $n_{(i)}$ is the number of times unit i is selected for the sample from any frame. Observe that $n_{(i)} = 0$ for the population units not in the sample. In the great majority of applications, $n_{(i)}$ will be one for most sampled units, but $n_{(i)} > 1$ is a possibility with this design.

The expected number of times unit i will be selected for the sample is $E[n_{(i)}] = \sum_f p_{if}$, where p_{if} is the probability of selecting unit i in the stratified simple random sample from frame f ; that is, $p_{if} = n_{fh} / N_{fh}$, where unit i is in stratum h of frame f .

There is also a Horvitz-Thompson estimator for T under the design, namely $t_{HT} = \sum_{i \in S} y_i / \pi_i$, where S denotes the sample and $\pi_i = 1 - (1 - p_{i1})(1 - p_{i2}) \cdots (1 - p_{iF})$. See Bankier (1986) for further discussion of this approach.

¹ Phillip S. Kott, Research Division; John F. Amrhein, Survey Sampling Branch; and Susan D. Hicks, Estimates Division, National Agricultural Statistics Service, USDA.

3. SAMPLING STRATEGIES USING PERMANENT RANDOM NUMBERS

The sampling design discussed above is independent across frames. For many surveys, however, it would be convenient if the design were *not* independent across frames. This is because all units in the combined sample are given the same survey instrument, and many units are in a number of frames. Therefore, given frame/stratum sample-size targets, a design with a tendency towards selecting the same unit in every frame should result in a smaller overall number of contacts (and consequently survey costs) than independent sampling across frames.

To this end, suppose each unit has been given a target p_{if} in each frame to meet or exceed. This target value is constant for all units in stratum h of frame f . We will withhold judgement on the policy of focussing on target p_{if} values – or equivalently on target n_{if} values – until the concluding section. Suffice it to say that many statistical agencies, including NASS, have such a policy.

One potential sampling design assigns each unit in the population a *permanent random number* (PRN) drawn from the uniform distribution on the interval $[0, 1)$. Unit i is selected for the frame f sample when its PRN is less than p_{if} .

The result is a Poisson sample where the probability of selecting unit i for the sample is $\pi_i = \max_f \{p_{if}\}$, which is clearly at least as large as each individual p_{if} for a given unit. An unbiased Horvitz-Thompson estimator for T under this design is $t_p = \sum_{i \in S} y_i / \max_f \{p_{if}\}$.

Under Poisson sampling, sample size is random. One way to reduce the variance of the sample size is with a variant of this sample design. In *collocated* PRN sampling, each population unit is assigned a unique PRN from among the members of the set $\{e/N, (1+e)/N, (2+e)/N, \dots, (N-1+e)/N\}$, where e is a uniform random variable drawn from the interval $[0, 1)$. To this end, one can first draw provisional PRN's for each unit followed by a value for e . The unit with the smallest provisional PRN is assigned a collocated PRN of e/N , the units with the second smallest provisional PRN is assigned $(1+e)/N$, and so on until $(N-1+e)/N$ is assigned to the unit with the largest provisional PRN. The estimator t_p remains unbiased under collocated sampling.

Due to random nature of the sample sizes resulting from Poisson and collocated sampling, frame/stratum sample size targets may not be met when a particular sample is drawn. A third PRN design begins with target n_{fh} values and removes this possibility. In this design, the units in stratum h of frame f with the n_{fh} smallest PRN's are selected for the sample (this is very similar to sequential Poisson sampling in Ohlsson 1995). A Horvitz-Thompson estimator under this *fixed-sample-size* PRN design requires one to compute the selection probabilities of the sampled units – a difficult task which may have to be approximated by simulation.

4. A SYSTEMATIC PROBABILITY PROPORTIONAL TO SIZE DESIGN

Another sampling design with the same selection probabilities as the Poisson (and collocated) sampling scheme described in the last section consists of the following steps:

- 0) When necessary, create an additional “stratum” for each frame consisting of those units not in any design stratum.
- 1) Divide up the population into mutually exclusive cells by cross-classifying the strata from the various frames. A pair of units in a particular cell will then be in the same stratum of each frame (e.g., the large oats stratum, the medium grain stocks stratum, and the no sorghum stratum).
- 2) Randomly order the population units in each cell and then sort the cells themselves in any order. This results in a list of all population units.
- 3) Draw a systematic probability proportional to “size” (PPS) sample from this list using the π_i described in the discussion of Poisson sampling as the measures of size (the word “size” is in quotes because the π_i are not really size measures in a conventional sense). This ensures that a unit's selection probability equals π_i .

The systematic PPS sampling design introduced above will always result in a sample of size close to $\sum_{i \in P} \pi_i$. In fact, if $\sum_{i \in P} \pi_i$ is an integer, then the sample size will exactly equal that sum. Otherwise, the sample size will be one of the two integers closest to $\sum_{i \in P} \pi_i$. Similarly, the expected number of sampled units in a cell, C , will be $\sum_{i \in C} \pi_i$, while the actual sample size will either be $\sum_{i \in C} \pi_i$ or one of the two integers closest to it.

Consider now a particular stratum h in a particular frame f with target sample size n_{fh} . For a unit i in this stratum, $\pi_i \geq n_{fh}/N_{fh}$ by design. Let $P(fh)$ denote the set of population units in stratum fh . The expected number of sampled units in fh is $\sum_{i \in P(fh)} \pi_i \geq n_{fh}$. There is no *guarantee* that the realized sample size in the stratum will be greater than or equal to n_{fh} . Nevertheless, given the above inequality and the lower bounds on the sample sizes of the cells within fh , the sample size in stratum fh will never be far below n_{fh} .

The advantages of this design over Poisson and collocated sampling is that it produces a more stable size and a greater likelihood of meeting frame/stratum requirements. Fixed-sample-size PRN, by contrast, will always meet frame/stratum requirements, but it does so at a cost: the design has a less stable overall sample size, and selection probabilities can be very difficult to determine.

5. EVALUATION OF THE ALTERNATIVE SAMPLING TECHNIQUES

To evaluate these sampling techniques empirically, we selected three states that conduct NASS's Vegetable Chemical Use Survey and replicated the three PRN techniques, the systematic PPS method, and independent sampling across frames 100 times. The assigned PRN's were maintained across the three PRN techniques within each replicate. A separate frame was constructed for each commodity of interest within a state (the number of frames ranged from two in Minnesota to 23 in California). Population units were allocated to one of four strata in each frame; two probability strata, one take-all stratum, and one zero stratum were used in each frame. Stratum boundaries were determined using a modified Lavallée and Hidioglou (1988) method, and units were assigned to strata based on a $\text{cum}^3\sqrt{f(x)}$ rule (Sweet and Sigman 1995). This stratification was chosen to mimic what might be a reasonable or reasonably common univariate sample design.

A target sample size of one-third the population was selected from each of the probability strata. Table 1 compares the overall sample sizes realized from each of the sampling techniques. As expected, the independent frame approach realized the largest sample sizes. The three PRN techniques realized sample sizes of similar size with the Poisson method experiencing the highest standard deviations in each of 3 trials (states). The PPS method appears to be the most stable.

Table 1
Mean realized sample sizes (over 100 replications)

State	Independent Frame Method	Fixed Sample Size Method	Poisson PRN Method	Collocated PRN Method	Systematic PPS Method
CA	496 (8.8)	388 (9.6)	375 (11.1)	374 (5.6)	373 (.14)
MI	658 (9.3)	513 (9.2)	504 (13.6)	501 (6.0)	502 (.48)
NJ	563 (8.1)	359 (8.6)	343 (13.8)	344 (4.6)	343 (.17)

Population sizes are: CA-775; MI-1041; NJ-785.
Standard deviations are in parentheses.

Table 2 shows the percentage of strata-level Poisson and PPS samples that fell short of their target sample sizes. One reason more shortfalls were not observed in the Poisson methods' realized sample sizes is the occurrence of what we call "visitors". A visitor is a sample unit that was not chosen within a specific commodity's frame, but ends up in the sample because it was selected in another commodity's frame. The existence of visitors tend to cause frame-level sample sizes to be larger, on average, than the targeted sizes.

Figure 1 shows cumulative distributions of differences between realized and desired sample sizes as percents of the desired sample sizes for the sampled strata. That is, the cumulative distribution of (realized - desired)/desired at the probability stratum level. For example, Michigan had 13 commodity frames each with two probability strata. Sampling from these frames was replicated 100 times so that the cumulative distribution function (CDF) for each technique utilized 2600 points. The two Poisson methods are shown as a single line since they coincide. The Poisson methods do not over-sample as much as the fixed-sample-size and independent frame methods, but at the risk of under-sampling as we saw in Table 2. The fixed-sample-size techniques (with dependent and independent frames) do not experience under-sampling, but do experience more over-sampling than the Poisson and PPS methods. The PPS method experiences some under-sampling but not to the extent of the Poisson methods. The PPS design also shows the steepest gradient of all the CDF's, indicating that it realizes less over-sampling.

Table 2
Percentage of probability strata for which the realized sample size fell short of the target (in 100 replications)

State	Poisson PRN Method	Collocated PRN Method	Systematic PPS Method
CA	11%	11%	6.3%
MI	12%	12%	6.3%
NJ	11%	8%	1.4%

Under the Poisson and collocated techniques, the probability of selection for unit i is $\pi_i = \max_f(p_{fh})$ where h corresponds to the stratum in which i belongs for frame f . The same probability of selection is used for the PPS technique. By contrast, the probabilities of selection under the fixed-sample-size PRN method are difficult to determine and may need to be simulated.

Such a simulation was conducted using the California data. The fixed-sample-size technique was run 10,000 times. Since all probability strata were sampled at a rate of 1/3, the simulated probabilities (*i.e.*, relative frequencies) can be compared to 1/3. The mean simulated probabilities of selection over the 10,000 trials are shown in Figure 2 as a function of the number of frames in which the unit is contained within a probability stratum. There were 19 commodities of interest in this state, but no units existed in probability strata in exactly 16 or 19 frames. A unit's probability of selection tends to increase with the number probability strata containing it. This selection probability is 1/3 only when the unit is in exactly one such stratum.

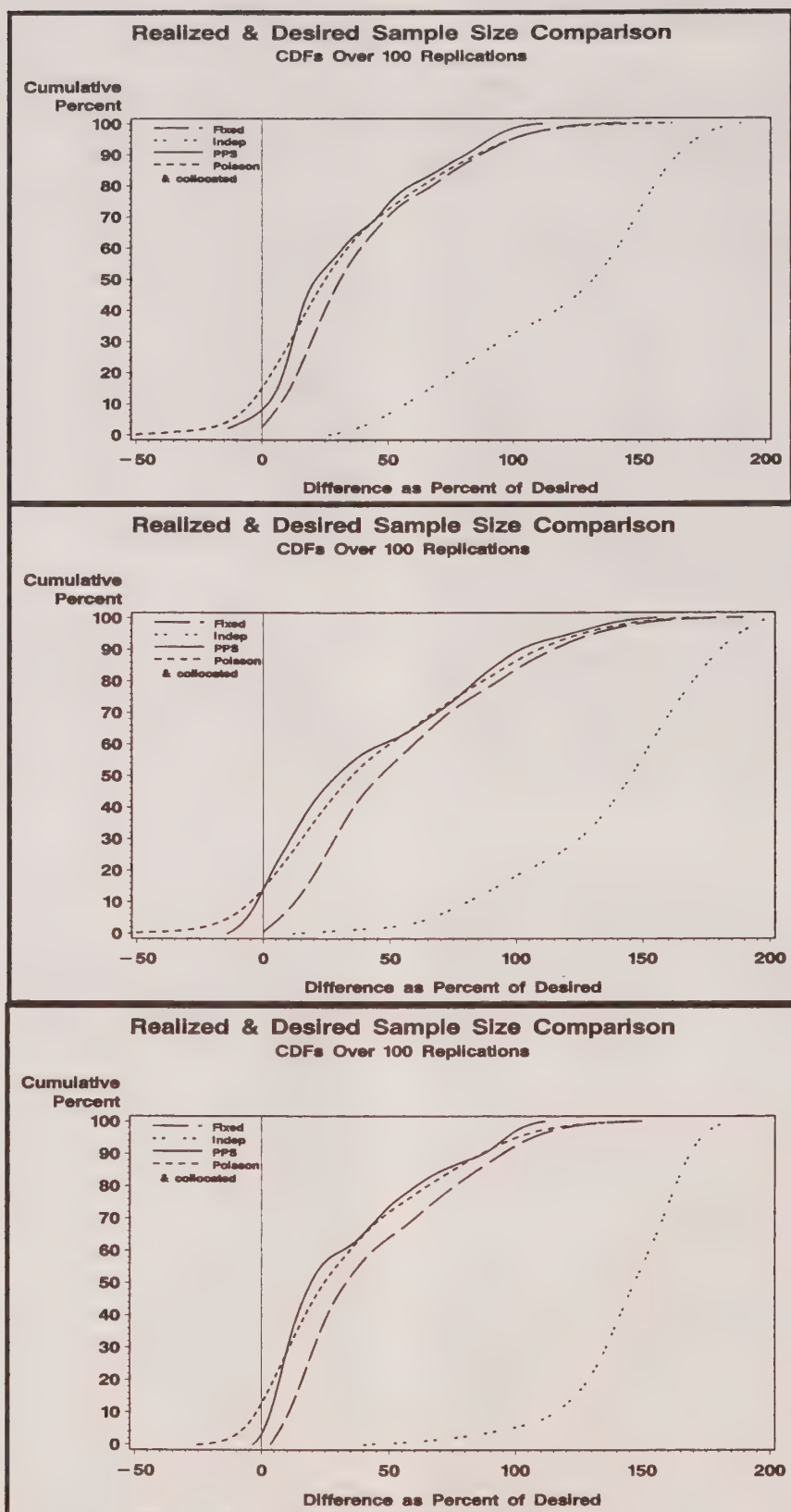


Figure 1. Comparison of realized and desired sample sizes for sampled strata. Top - MI; middle - CA; bottom - NJ.

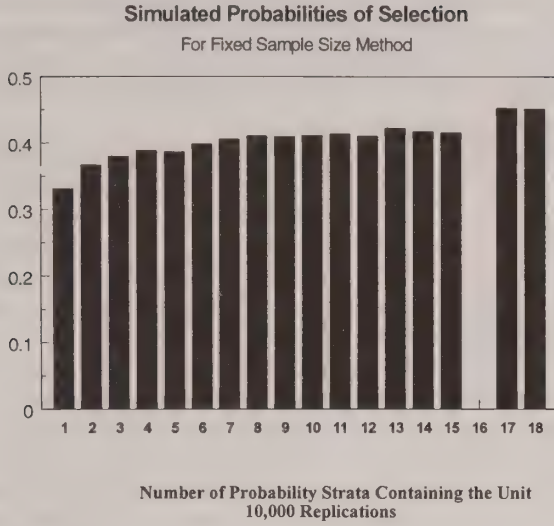


Figure 2. Simulated probabilities of selection for the fixed-sample-size method-California

6. CALIBRATION

The problem with both t_M and t_P (or t_{HT}) is that they are often not very good estimators for T in term of precision (variance). One of the properties of single-frame, stratified simple random sampling is that the conventional expansion estimator estimates the stratum population size perfectly (*i.e.*, with zero variance). In our multiple frame set up, however, neither t_M nor t_P will estimate the N_{fh} perfectly in most applications.

Let us define $w_i^0 = n_{(i)}/E[n_{(i)}]$ as the *original sampling weight* of unit i in t_M . Similarly, $w_i^0 = 1/\max_f \{p_{if}\}$ in t_P and $1/\pi_i$ more generally for a Horvitz-Thompson estimator. Bankier (1986) proposed raking to create a set of adjusted weights such that

$$\sum_{i \in S_{fh}} w_i^C = N_{fh} \quad (2)$$

for each stratum h in every frame f , where S_{fh} is that part of the sample that is in stratum h of frame f regardless of the frame(s) from which the units were selected.

Deville and Särndal (1992) call (2) a *calibration equation*. They point out that there are a number of ways to compute the *calibration weights*, the w_i^C , so that equation (2) is satisfied and w_i^C/w_i^0 is in some sense close to 1 for all i . One method is raking as suggested by Bankier (1986). Another method, discussed at length by Deville and Särndal (1992), uses least squares. Either way, the resulting estimator

$$t_C = \sum_{i \in S} w_i^C y_i,$$

where S denotes the entire sample, will be nearly design unbiased because w_i^C/w_i^0 is close to 1 for all i .

The estimator t_C is also unbiased under the model:

$$y_i = \beta_0 + \sum_{f=1}^F \sum_{h=2}^{H_f} d_{ifh} \beta_{fh} + \epsilon_i, \quad (3)$$

where the dummy variable, d_{ifh} , is 1 when unit i is in stratum h of frame f (sampled or not) and zero otherwise, while ϵ_i is a random variable with a mean of zero. The β_0 and the β_{fh} are unknown constants (β_0 represents the mean y -value for a unit in the first stratum of every frame; that is why the second sum excludes $h = 1$). The same d_{ifh} values apply to every survey item (y) of interest, while the β values change with the survey item. For many survey items, β_{fh} values will be zero when frame f (say, grain stocks) is irrelevant to the item (say, planted oat acres).

Isaki and Fuller (1982) call the model expectation of the design mean squared error of t_C the “anticipated mean squared error” of the estimator. This value is of most use at the planning stage of a sample survey.

If the model in equation (3) holds, and the ϵ_i are uncorrelated, then the anticipated mean squared error of t_C is

$$\begin{aligned} E_\epsilon[\text{MSE}_D(t_C)] &= E_\epsilon\{E_D[\sum_s w_i^C y_i - \sum_P y_i]^2\} \\ &= E_D\{E_\epsilon[(\sum_s w_i^C y_i - \sum_P y_i)^2]\} \\ &= E_D\{E_\epsilon[(\sum_s w_i^C \epsilon_i - \sum_P \epsilon_i)^2]\} \\ &= E_D\{\sum_s [(w_i^C)^2 - 2w_i^C]E_\epsilon(\epsilon_i^2) + \sum_P E_\epsilon(\epsilon_i^2)\} \\ &\approx E_D\{\sum_s [(1/\pi_i)^2 - 2/\pi_i]E_\epsilon(\epsilon_i^2) + \sum_P E_\epsilon(\epsilon_i^2)\} \\ &= \sum_P (1/\pi_i - 1)E_\epsilon(\epsilon_i^2), \end{aligned} \quad (4)$$

since $w_i^C \approx 1/\pi_i$. It is of some interest to note that using Poisson, collocated, and systematic PPS sampling result in estimators with approximately equal anticipated mean squared errors asymptotically. This surprising result is in part due to the nature of a calibrated estimator, but it is also a repercussion of the fact that when we take the design expectation of the approximate model variance in the last line of equation (4), we average over all possible samples and remove the biggest source of variation among the three sampling designs.

Now suppose we had used stratified simple random sampling and selected unit i with probability $p_{if} \leq \pi_i$, where f is the frame relevant to y . It is not hard to show that the anticipated variance of the simple expansion estimator would have been $\sum_P (1/p_{if} - 1)E_\epsilon(\epsilon_i^2)$, which is at least as large as the right hand side of equation (4). Thus, there are gains – in large samples, at least – from “integrating” the samples from various frames as we have effectively done. How large the samples must be in practice for the asymptotic results to be relevant is unclear. At the very least, the sample size must be many times the number of model parameters in equation (3).

A few words on mean squared error estimation for t_C are in order. The mean squared error estimator advocated by Deville and Särndal (1992) – an estimator with both good design and model-based properties – can not be implemented

unless the joint selection probability (π_{ij}) for every pair of sample units (i and j) is known. Among the designs we have discussed, these probabilities are easily calculated only for the Poisson variant of PRN (where $\pi_{ij} = \pi_i \pi_j$).

As we have observed in equation (4), the anticipated mean squared error of the calibration estimator is the same under Poisson PRN, collocated PRN, and systematic PPS sampling. This suggests that the Poisson mean squared error estimator may be reasonable under each of the three designs. A stronger model-driven argument exists for this contention, but will not be made here.

7. DISCUSSION

In the last section, it was pointed out that if calibration weights were designed to satisfy equation (2), the resulting estimator would be unbiased under the model in equation (3). In many applications, there may be a more appropriate model on which to base calibration than the one in equation (3). For example, if there was a continuous control variable used to stratify a particular frame, it makes more sense to use that variable directly in the model rather than indirectly through frame/stratum identifiers.

Raking is a form of calibration under a particular model. With that in mind, it makes sense to use the most reasonable model available. Least squares has the advantage over raking that it can easily be applied to continuous control variables. Singh and Mohl (1996) provide an extensive review of alternative calibration algorithms including an extension of raking to continuous variables. An intriguing least-squares variant missed by Singh and Mohl (1996) can be found in Brewer (1994).

Many economic and agricultural surveys employ rotating sample designs. This has proved an effective way to balance cost and burden considerations. Although our empirical findings demonstrated an advantage of the systematic PPS methodology in terms of meeting target sample sizes, the three PRN designs are much more conducive to sample rotation. See, for example, Ohlsson (1995) on this topic. Moreover, with the PRN methods, one can integrate different frames at different times of the year (with systematic PPS there is no easy way to allocate the sample back to the frame of origin). This is a particularly useful property for agricultural surveys because different crops have different growing seasons.

In summary, the fixed-sample-size PRN sample design is excellent for meeting target sample sizes but is hard to use in practice because selection probabilities are usually unknown and must be simulated. The systematic PPS design is very good at meeting target sample sizes but is difficult to incorporate into a sample rotation scheme. Moreover, mean squared error estimation requires invocation of model assumptions. Our empirical example shows that collocated sampling may only be slightly better than Poisson at meeting target sample sizes. It should be recognized, however, that other configurations of the frames,

strata, and sampling fractions may produce different results. Moreover, collocated sampling is conducive to rotation schemes, like Poisson sampling. On the other hand, like PPS sampling, it requires the assumption of a model to estimate mean squared error.

Finally, setting p_{if} or n_{if} targets is a popular, but indirect, means of controlling the variance of the estimator t_C associated with each frame. These targets lead to our *ad hoc* decision to set π_i equal to $\max_f \{p_{if}\}$. A more direct strategy would be to set (asymptotic) anticipated variance targets for each frame estimator using equation (4) and postulated values for the $E_{\epsilon} (\epsilon_i^2)$. One could then choose, say, the set of π_i that minimizes the expected sample size yet satisfy these variance targets. A similar approach is taken by Amrhein, Fleming, and Bailey (1997) who use Chromy's algorithm in a manner analogous to Sigman and Monsour (1995). Poisson PRN, collocated PRN, and systematic PPS sampling remain three viable alternatives for selecting the sample once optimal π_i are determined.

REFERENCES

- AMRHEIN, J.F., FLEMING, C.M., and BAILEY, J.T. (1997). Determining the probabilities of selection in a multivariate probability proportional to size sample design. In *Proceedings: Symposium 97: New Directions in Surveys and Censuses*. Statistics Canada. To appear.
- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- BREWER, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10(1)A, 213-233.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimator in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTT, P.S., and FETTER, M.J. (1997). A multi-phase sample design to co-ordinate surveys and limit response burden. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. To appear.
- LAVALLÉE, P., and HIDIROGLOU, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33-43.
- OHLSSON, E. (1995). Coordination of samples using permanent random numbers. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: Wiley, 153-169.
- SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimators in survey sampling. *Survey Methodology*, 22, 107-115.

- SIGMAN, R.S., and MONSOUR, N.J. (1995). Selecting samples from list frames of businesses. In *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, and P.S. Kott). New York: Wiley, 133-152.
- SKINNER, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- SKINNER, C.J., HOLMES, D.J., and HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 3, 333-347.
- SWEET, E., and SIGMAN, R.S. (1995). *User Guide for the Generalized SAS Univariate Stratification Program*, Economical Statistical Methods and Programming Division, Bureau of the Census, U.S. Department of Commerce, Report number ESM-9504.

Use of Auxiliary Information for Two-phase Sampling

M.A. HIDIROGLOU and C.-E. SÄRNDAL¹

ABSTRACT

Two-phase sampling designs offer a variety of possibilities for use of auxiliary information. We begin by reviewing the different forms that auxiliary information may take in two-phase surveys. We then set up the procedure by which this information is transformed into calibrated weights, which we use to construct efficient estimators of a population total. The calibration is done in two steps: (i) at the population level; (ii) at the level of the first-phase sample. We go on to show that the resulting calibration estimators are also derivable via regression fitting in two steps. We examine these estimators for a special case of interest, namely, when auxiliary information is available for population subgroups called calibration groups. Poststrata are the simplest example of such groups. Estimation for domains of interest and variance estimation are also discussed. These results are illustrated by applying them to two important two-phase designs at Statistics Canada. The general theory for using auxiliary information in two-phase sampling is being incorporated into Statistics Canada's Generalized Estimation System.

KEY WORDS: Generalized regression; Two-phase sampling; Model assisted approach; Domain estimation; Calibration factors.

1. INTRODUCTION

Two-phase sampling is a powerful and cost-effective technique. It was first proposed by Neyman (1938). In Cochran's (1977) book, and in its two earlier editions dated 1953 and 1963, one finds basic results for two-phase sampling, including the simplest regression estimators for such designs. This paper takes a broader outlook and proposes a general approach to the use of auxiliary information in two-phase survey designs. Our main references are Särndal and Swensson (1987), Särndal, Swensson and Wretman (1992) and Dupont (1995). Recent related work includes Breidt and Fuller (1993), who presented computationally efficient estimation procedures for three-phase sampling in the presence of auxiliary information. Chaudhuri and Roy (1994) studied optimality properties of the well-known simpler regression estimators for two-phase sampling. Binder (1996) described a simple linearization procedure to estimate variances of nonlinear estimators. His procedure can be applied to any sampling design, including two-phase-sampling. Throughout this paper, we assume *arbitrary* sampling designs for each of the two phases.

Single-phase sampling involves the use of one layer of information for estimation. In two-phase sampling, however, one has to consider two layers of information. This complicates matters, and it is not clear-cut how best to exploit the combined information from the two sources. Two approaches are considered in this paper for building estimators based on auxiliary information. These are the *calibration approach* and the *generalized regression approach*. We show that the generalized regression approach can be viewed as a special case of the calibration

approach. The two approaches are examined under a common structure for the auxiliary information. It assumes that information exists about an auxiliary vector \mathbf{x}_1 for the units of the entire population, and about a second auxiliary vector \mathbf{x}_2 for the units of the first phase sample. Consequently, at the level of the first phase sample, there is information about both vectors, \mathbf{x}_1 and \mathbf{x}_2 .

The *generalized regression approach*, as applied to two-phase sampling, is discussed in Särndal *et al.* (1992). These authors develop the general regression estimator for two-phase sampling, assuming arbitrary sampling designs in each of the two phases. Two regression fits are carried out. A "bottom level" regression is fitted to produce predicted values up to the level of the first phase sample, using the auxiliary information available for this step. Next, a "top level" regression is fitted to produce predicted values up to the entire population level, using the information appropriate for this step. The two sets of predicted values are used to build a generalized regression estimator.

The *calibration approach* focuses on the weights given to the units for purposes of estimation. Calibration implies that a set of starting weights (usually the sampling design weights) are transformed into a set of new weights, called calibrated weights. The calibrated weight of a unit is the product of its initial weight and a calibration factor. The calibration factors are obtained by minimizing a function measuring the distance between the initial weights and the calibrated weights, subject to the constraint that the calibrated weights yield exact estimates of the known auxiliary population totals. In two-phase sampling the two levels of information imply two consecutive calibrations. The first phase of calibration uses the auxiliary information available (at least population counts) at the level of the entire

¹ M.A. Hidiroglou, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, K1A 0T6; and C.-E. Särndal, University of Montreal, and Statistics Canada.

population, resulting in first-phase calibrated weights. The second phase of calibration uses these first-phase calibrated weights and incorporates the information at the level of the first-phase sample, resulting in a final set of calibrated weights.

Both approaches profit from the two layers of information. They do not necessarily yield identical results. Whether they do or not depends on the exact formulations given to the regression fits and the calibration approach. This is apparent in Dupont (1995), where four alternative estimators were developed under the regression approach. These differ in the way that the auxiliary variables are used in deriving the predicted y -values required for the regression estimator. For each of these four approaches, Dupont built a matching estimator using the calibration approach. She succeeded in obtaining an exact equivalence between the two approaches in only one of the four cases. Three of Dupont's four approaches can be considered as special cases of the general approach in this paper.

In this paper, building on Hidirolou and Särndal (1995), we provide a unified theory for two-phase sampling with auxiliary information. We show that the regression estimators can be obtained as a special case of the calibration approach. Direct linkage between the two approaches is therefore possible. One motivation for our work was the necessity to provide tools for efficient use of administrative data sources in several important Statistics Canada surveys. Our work has prepared the way for the inclusion of two-phase sampling into Statistics Canada's Generalized Estimation System described in Estevao, Hidirolou and Särndal (1995).

We illustrate our general theory by applying it to two survey designs currently used at Statistics Canada. The first application, Armstrong and St-Jean (1994), describes the use of the two-phase approach for sampling tax records. Our second application, Hidirolou, Latouche, Armstrong and Gossen (1995), involves the use of two-phase sampling of payroll deduction accounts used in Statistics Canada's Survey of Earnings, Payrolls and Hours.

The paper is organized as follows. Section 2 sets up the notation. Section 3 specifies our version of the calibration approach in two-phase sampling. Section 4 establishes the important result that the resulting calibration estimator can be expressed, with exact equivalence, as a two-phase regression estimator, that is, one derived via two consecutive regression fits. Additional theoretical results are reported in Sections 5 and 6. Section 5 examines the forms taken by our two-phase calibration estimator under important special types of information, namely, when some of the auxiliary variables, either in the first or in the second phase, correspond to categorical variables that codify a grouping of the units into mutually exclusive and exhaustive classes. Section 6 gives results on two issues that always require attention in a survey, which are central to the GES, namely, (a) estimation for domains (sub-populations), and (b) design-based variance estimation. For variance estimation

we use the approach of Särndal and Swensson (1987). Section 7 shows how the preceding theory is applied to two-phase designs currently in use at Statistics Canada. Finally, Section 8 provides a brief summary.

2. NOTATION

The population is represented by $U = \{1, \dots, k, \dots, N\}$. A first-phase probability sample $s_1 (s_1 \subseteq U)$ is drawn from the population U , according to a sampling design with the selection probabilities $\pi_{1k} = P(k \in s_1)$. Given s_1 , a second-phase sample $s_2 (s_2 \subseteq s_1 \subseteq U)$ is selected from s_1 , according to a sampling design with the selection probabilities $\pi_{2k} = P(k \in s_2 | s_1)$. Note that these are conditional probabilities, given s_1 . We assume that $\pi_{1k} > 0$ for all $k \in U$ and $\pi_{2k} > 0$ for all $k \in s_1$. From this point on, we work with weights in the estimation process. We will denote the first-phase sampling weight of unit k as $w_{1k} = 1/\pi_{1k}$, and the second-phase sampling weight as $w_{2k} = 1/\pi_{2k}$. The overall sampling weight for a selected unit is $w_k^* = w_{1k} w_{2k}$.

Our objective is to estimate the population total $Y = \sum_U y_k$, where y_k is the value of the variable of interest y for unit k . If $A \subseteq U$ is an arbitrary set of units, we write simply \sum_A for $\sum_{k \in A}$. The customary two-phase sampling procedure calls for collecting inexpensive information about the units k belonging to a large first-phase sample s_1 . This information is then used to realize efficient sampling and estimation in the second phase. The values y_k are recorded for $k \in s_2$. An unbiased estimator of Y is given by $\hat{Y} = \sum_{s_2} w_k^* y_k$. This estimator uses sampling weights only. A more extensive use of available auxiliary information is achieved through the regression estimators that we will now examine.

We denote the auxiliary vector at the level of the first-phase sample as \mathbf{x} and its value for unit k as \mathbf{x}_k . As in Särndal *et al.* (1992, chapter 9), we partition \mathbf{x}_k as $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$. Information is available up to the entire population level for the vector \mathbf{x}_{1k} , whereas for the vector \mathbf{x}_{2k} , information is only available up to the level of the first-phase sample. Table 1 summarizes our assumptions on the auxiliary information available for estimation.

Table 1
Relationship between set of units and available data
at different levels

Set of units	Data available
Population	$\{\mathbf{x}_{1k} : k \in U\}$ or $\sum_U \mathbf{x}_{1k}$
First-phase sample	$\{\mathbf{x}_k : k \in s_1\}$
Second-phase sample	$\{(\mathbf{x}_k, y_k) : k \in s_2\}$

Note that individual values \mathbf{x}_{1k} , $k \in U$, are not required. It suffices to know the total $\sum_U \mathbf{x}_{1k}$, which may be taken from a reliable administrative source. The presence of auxiliary information in one or both phases opens the possibility of modifying the sampling weights with the aid of calibration

factors calculated using the auxiliary information. In each of the two phases, a unit's sampling weight is modified by multiplying it by the calibration factor, resulting in a calibrated weight.

The first-phase calibrated weight \tilde{w}_{1k} is computed for units $k \in s_1$ as $\tilde{w}_{1k} = w_{1k} g_{1k}$. The first-phase sampling weight is w_{1k} , and the first-phase calibration factor is g_{1k} . Similarly, we compute overall calibration weights $\tilde{w}_k^* = w_k^* g_k^*$ for units $k \in s_2$, where g_k^* is the overall calibration factor. The superscript "*" denotes overall weights taking into account both phases. The superimposed symbol "~" denotes calibrated weights.

3. CALIBRATION WITH GENERALIZED LEAST SQUARES DISTANCE

Auxiliary information available at each phase of sampling can improve weights by the process known as calibration. This improvement yields smaller variances of the resulting estimates if there is a strong correlation between the auxiliary variables and the variables of interest. We seek a set of "new" weights that lie as close as possible to a set of starting weights. The calibration requires the specification of a measure of the distance between the starting weights and the new weights. Several distance functions have been proposed; see Deville and Särndal (1992), Deville, Särndal, and Sautory (1993), and Singh and Mohl (1996). Any one of these distance functions could be used for two-phase calibration. However, we concentrate on one of these, namely, the generalized least squares (GLS). For an arbitrary set of units s , it is of the form

$$D = \frac{1}{2} \sum_s C_k \frac{(\tilde{w}_k - w_k)^2}{w_k} \quad (3.1)$$

where $\{w_k : k \in s\}$ are the starting weights, $\{\tilde{w}_k : k \in s\}$ are the new calibrated weights, and $\{C_k : k \in s\}$ are specified positive factors that control the relative importance of the terms of this sum. For each of the two phases, we minimize a GLS distance measure with suitable factors C_k , subject to constraints. After applying the two successive calibrations, we have a set of overall calibrated weights.

(i) First-phase calibration (from s_1 to U).

The first-phase sampling weights $\{w_{1k} : k \in s_1\}$ are used as starting weights. Let $\{C_{1k} : k \in s_1\}$ be pre-specified positive factors. We determine the first-phase calibrated weights by minimizing the GLS distance

$$D_1 = \frac{1}{2} \sum_{s_1} C_{1k} \frac{(\tilde{w}_{1k} - w_{1k})^2}{w_{1k}} \quad (3.2)$$

subject to the first-phase calibration equation

$$\sum_{s_1} \tilde{w}_{1k} x_{1k} = \sum_U x_{1k} \quad (3.3)$$

where the total $\sum_U x_{1k}$ is known. Note that this calibration does not involve information concerning x_{2k} because it is available only up to s_1 .

The resulting weights are

$$\tilde{w}_{1k} = w_{1k} g_{1k} \quad (3.4)$$

with

$$g_{1k} = 1 + \left(\sum_U x_{1k} - \sum_{s_1} w_{1k} x_{1k} \right)' T_1^{-1} \frac{x_{1k}}{C_{1k}} \quad (3.5)$$

and

$$T_1 = \sum_{s_1} \frac{w_{1k} x_{1k} x_{1k}'}{C_{1k}} \quad (3.6)$$

Some of the \tilde{w}_{1k} given by (3.4) may be negative, or zero. Many users prefer weights to be always positive. This can be achieved by adding to (3.3) the inequality constraints $\tilde{w}_{1k} > 0$ for all $k \in s_1$. The resulting weights have no closed expression, in contrast to (3.4).

(ii) Second-phase calibration (from s_2 to s_1).

We use $\{\tilde{w}_{1k} w_{2k}, k \in s_2\}$ as starting weights, where \tilde{w}_{1k} is given by (3.4). These weights incorporate the information about x_{1k} available up to the full population level. Applying them to the data $\{y_k : k \in s_2\}$ yields one possible estimator, namely $\hat{Y} = \sum_{s_2} \tilde{w}_{1k} w_{2k} y_k$. However, since these weights do not contain the x_{2k} -value information available for $k \in s_1$, they can be improved through a second-phase calibration. Let $\{C_{2k} : k \in s_2\}$ be specified positive factors. We determine the overall calibrated weights \tilde{w}_k^* by minimizing

$$D_2 = \frac{1}{2} \sum_{s_2} \frac{C_{2k} (\tilde{w}_k^* - \tilde{w}_{1k} w_{2k})^2}{\tilde{w}_{1k} w_{2k}} \quad (3.7)$$

subject to the second-phase calibration equation

$$\sum_{s_2} \tilde{w}_k^* x_k = \sum_{s_1} \tilde{w}_{1k} x_k \quad (3.8)$$

where $x_k = (x_{1k}', x_{2k}')'$. The resulting overall calibrated weights are

$$\tilde{w}_k^* = w_k^* g_k^* \quad (3.9)$$

where

$$g_k^* = g_{1k} g_{2k} \quad (3.10)$$

with g_{1k} given by (3.5) and g_{2k} by

$$g_{2k} = 1 + \left(\sum_{s_1} \tilde{w}_{1k} x_k - \sum_{s_2} \tilde{w}_{1k} w_{2k} x_k \right)' T_2^{-1} \frac{x_k}{C_{2k}} \quad (3.11)$$

for $k \in s_2$, and

$$T_2 = \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \mathbf{x}_k \mathbf{x}'_k}{C_{2k}} \quad (3.12)$$

Again, some g_k^* may be zero or negative, but always positive g_k^* can be ascertained by adding to (3.8) the inequality constraints $w_k^* > 0$ for $k \in s_2$.

Having determined the overall weights \tilde{w}_k^* by equation (3.9), the estimator of Y is given by

$$\hat{Y} = \sum_{s_2} \tilde{w}_k^* y_k \quad (3.13)$$

Remark 3.1 A potential problem with the above approach is that some of the g_{1k} 's may be negative or even zero. If this occurs, (3.7) is not a proper distance measure. Some of the important applications, such as poststratification, do not have this problem as their associated g_{1k} 's are always greater than zero. If all the g_{1k} 's are greater than zero, then the minimization criterion given by (3.7) is acceptable. Otherwise, we have to modify it. One possible modification is to impose on the above-mentioned constraints that the w_{1k} 's are positive for $k \in s_1$. Another possible modification is to replace C_{2k} in (3.7) by

$$C_{2k}^* = C_{2k} \frac{\tilde{w}_{1k}}{w_{1k}}.$$

Then

$$\frac{C_{2k}^*}{\tilde{w}_{1k} w_{2k}} = \frac{C_{2k}}{w_k^*},$$

which is always positive. The resulting g_k^* -factors in (3.9) can be shown to be $g_k^* = g_{1k} + g_{2k} - 1$, where g_{1k} is given as before by (3.5), and g_{2k} by (3.11) provided that we instead define T_2 as

$$T_2 = \sum_{s_2} \frac{w_k^* \mathbf{x}_k \mathbf{x}'_k}{C_{2k}}.$$

It is our opinion that in most applications the choice between the multiplicative $g_k^* = g_{1k} g_{2k}$ and the additive form $g_k^* = g_{1k} + g_{2k} - 1$ would have little effect on the resulting estimates. That is, we believe the two point estimates would be very close, and so would be their associated estimates of variance.

Remark 3.2: Bounding the weights ordinarily has negligible impact on the estimates. Recent experience with calibration for single phase designs, Stukel, Hidirolou, and Särndal (1996), has shown that mildly different sets of g -weights lead to point estimates that differ very little. Some recently developed computer software for calibration, for example, the software described in Deville *et al.* (1993), minimizes a distance function such that the resulting

g -factors are guaranteed to be bounded from above and from below.

Remark 3.3: The auxiliary data in Table 1 can be used in several ways for two-phase calibration. Considering in particular the second-phase calibration equation defined by (3.8), three different specifications of the vector \mathbf{x}_k are: (i) $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$; (ii) $\mathbf{x}_k = \mathbf{x}_{2k}$; and (iii) $\mathbf{x}_k = \mathbf{x}_{1k}$. We comment on these possibilities, assuming for each of these that a first-phase calibration has been carried out, resulting in the first-phase calibrated weights (3.4).

The case (i) specification $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$, recommended in Särndal *et al.* (1992), capitalizes on all the available information. Thus, in this respect case (i) is ideal. Cases (ii) and (iii) disregard some available information. Case (ii) is sometimes of interest, despite some loss of information; an example is given in Section 7.1. Case (iii) implies that the data $\{\mathbf{x}_{2k} : k \in s_1\}$ are observed, but not used; we do not further consider this case. We call $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ the *full vector* and $\mathbf{x}_k = \mathbf{x}_{2k}$ the *reduced vector*.

Second-phase calibration on the reduced vector $\mathbf{x}_k = \mathbf{x}_{2k}$ can be carried out without significant loss of information if \mathbf{x}_{2k} is a good *substitute* for \mathbf{x}_{1k} , as also observed by Dupont (1995). However, if \mathbf{x}_{1k} complements \mathbf{x}_{2k} , then the full vector $\mathbf{x}_k = (\mathbf{x}'_{1k}, \mathbf{x}'_{2k})'$ should clearly be used in the calibration defined by (3.7). Otherwise, significant loss of information and increased variance may result.

Remark 3.4: Both the full and the reduced \mathbf{x}_k -vectors lead to overall weights \tilde{w}_k^* calibrated on \mathbf{x}_{2k} from s_2 to s_1 . This means that $\sum_{s_2} \tilde{w}_k^* \mathbf{x}_{2k} = \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{2k}$, because (3.8) holds, and \mathbf{x}_{2k} is contained in \mathbf{x}_k . However, there exists a difference between the full and reduced vector specifications with respect to the calibration on \mathbf{x}_{1k} . If the full vector specification is used in phase two, the resulting overall weights \tilde{w}_k^* are calibrated on \mathbf{x}_{1k} from s_2 to s_1 , and from s_1 to U . This means that $\sum_{s_2} \tilde{w}_k^* \mathbf{x}_{1k} = \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. In contrast, if the reduced vector specification is used, the resulting overall weights \tilde{w}_k^* are calibrated on \mathbf{x}_{1k} from s_1 to U by virtue of the first-phase calibration. That is $\sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. However, they are **not** calibrated from s_2 to s_1 , because \mathbf{x}_{1k} is not present in the second-phase calibration. Hence, $\sum_{s_2} \tilde{w}_k^* \mathbf{x}_{1k} \neq \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_U \mathbf{x}_{1k}$. Thus if the survey requires a weight system that will reproduce the known $\sum_U \mathbf{x}_{1k}$, then the full vector specification must be used.

So far, we have focused on the general framework for calibration with two levels of auxiliary information. This framework does not reveal the many interesting forms that the estimator \hat{Y} given by (3.13) may take for specific cases of auxiliary information. Some illustrations are given in Section 7. We first address three issues that are of practical interest in virtually every major survey: (i) poststratification or, more generally, the presence of auxiliary information for population subgroups (Section 5), (ii) estimation for domains of interest (Section 6), and (iii) the construction of variance estimates (Section 6).

4. THE TWO-PHASE CALIBRATION ESTIMATOR VIEWED AS A REGRESSION ESTIMATOR

An alternative expression for the calibration estimator (3.13) is given by formula (4.1) below. This expression links it exactly with the regression estimator for two-phase designs introduced in Särndal *et al.* (1992, chapter 9).

Theorem 4.1: When the overall calibrated weights \tilde{w}_k^* are determined by (3.9), the calibration estimator (3.13) is identical to the two-phase regression estimator given by

$$\hat{Y} = \sum_U \hat{y}_{1k} + \sum_{s_1} w_{1k} (\hat{y}_{2k} - \hat{y}_{1k}) + \sum_{s_2} w_k^* (y_k - \hat{y}_{2k}) \quad (4.1)$$

where \hat{y}_{1k} and \hat{y}_{2k} are successive regression predictions such that

$$\hat{y}_{1k} = \mathbf{x}_{1k}' \hat{\mathbf{B}}_1 \quad (4.2)$$

with

$$\hat{\mathbf{B}}_1 = \mathbf{T}_1^{-1} \left\{ \sum_{s_1} \frac{w_{1k} \mathbf{x}_{1k} \hat{y}_{2k}}{C_{1k}} + \sum_{s_2} \frac{w_k^* \mathbf{x}_{1k} (y_k - \hat{y}_{2k})}{C_{1k}} \right\} \quad (4.3)$$

where \mathbf{T}_1 is given by (3.6), and

$$\hat{y}_{2k} = \mathbf{x}_k' \hat{\mathbf{B}}_2 \quad (4.4)$$

with

$$\hat{\mathbf{B}}_2 = \mathbf{T}_2^{-1} \sum_{s_2} \frac{\tilde{w}_{1k} w_{2k} \mathbf{x}_k y_k}{C_{2k}} \quad (4.5)$$

where \mathbf{T}_2 is given by (3.12).

The proof for Theorem 4.1 uses some tedious but straightforward algebra and is not presented here.

We now show that (4.1) can be constructed via regression estimation in two steps. For the first step, suppose that the variable of interest y_k were observed for the full first-phase sample s_1 . The auxiliary information on \mathbf{x}_{1k} is available for $k \in s_1$ and the population total $\sum_U \mathbf{x}_{1k}$ is known. The resulting regression estimator of $Y = \sum_U y_k$ would then be given by

$$\begin{aligned} \hat{Y} &= \sum_U \hat{y}_{1k}^0 + \sum_{s_1} w_{1k} (y_k - \hat{y}_{1k}^0) \\ &= \sum_{s_1} w_{1k} y_k + \left(\sum_U \hat{y}_{1k}^0 - \sum_{s_1} w_{1k} \hat{y}_{1k}^0 \right) \end{aligned} \quad (4.6)$$

In the last expression, the first term represents the (hypothetical) first-phase Horvitz-Thompson estimator of Y . The second and third terms represent a regression adjustment, where \hat{y}_{1k}^0 is the predictor of y_k based on the fitted regression of y_k on \mathbf{x}_{1k} for $k \in s_1$. That is, $\hat{y}_{1k}^0 = \mathbf{x}_{1k}' \hat{\mathbf{B}}_1^0$, with

$$\hat{\mathbf{B}}_1^0 = \mathbf{T}_1^{-1} \sum_{s_1} \frac{w_{1k} \mathbf{x}_{1k} y_k}{C_{1k}}.$$

Note that $\sum_U \hat{y}_{1k}^0 = (\sum_U \mathbf{x}_{1k})' \hat{\mathbf{B}}_1^0$ where $\sum_U \mathbf{x}_{1k}$ is known. However, none of the terms in (4.6) can be computed directly, because y_k is only observed for the second-phase sample. A second step of regression estimation is thus necessary. It is carried out by replacing the unknown $\sum_{s_1} w_{1k} y_k$ in (4.6) by its conditional regression estimator

$$\sum_{s_1} w_{1k} \hat{y}_{2k} + \sum_{s_2} w_k^* (y_k - \hat{y}_{2k}) \quad (4.7)$$

where $\hat{y}_{2k} = \mathbf{x}_k' \hat{\mathbf{B}}_2$, with $\hat{\mathbf{B}}_2$ given by (4.5), is the predictor of y_k based on the regression of y_k on \mathbf{x}_k , known up to s_1 . Next, the vector $\hat{\mathbf{B}}_1^0$ required for computing \hat{y}_{1k}^0 contains a known matrix \mathbf{T}_1 and an unknown vector

$$\sum_{s_1} \frac{w_{1k} \mathbf{x}_{1k} y_k}{C_{1k}}.$$

Using a regression estimator for this unknown vector, we obtain $\hat{\mathbf{B}}_1$ given by (4.3) as a replacement for $\hat{\mathbf{B}}_1^0$. These two substitutions in (4.6) lead to the two-phase regression estimator given by (4.1), which is identical to the calibration estimator (3.13).

Remark 4.1: A more direct alternative to $\hat{\mathbf{B}}_1$ in (4.3) would be to use only the second-phase sample. This would have produced

$$\hat{\mathbf{B}}_{1,alt} = \left(\sum_{s_2} \frac{w_k^* \mathbf{x}_{1k} \mathbf{x}_{1k}'}{C_{2k}} \right)^{-1} \sum_{s_2} \frac{w_k^* \mathbf{x}_{1k} y_k}{C_{2k}}$$

The resulting predictions $\hat{y}_{1k,alt} = \mathbf{x}_{1k}' \hat{\mathbf{B}}_{1,alt}$ would be replacing \hat{y}_{1k} in (4.1). However, the resulting regression estimator is not identical to (3.13) and is a less efficient alternative, because $\hat{\mathbf{B}}_{1,alt}$ uses less \mathbf{x}_{1k} -information than $\hat{\mathbf{B}}_1$.

5. CALIBRATION GROUPS

In this Section we apply the results of Sections 3 and 4 to the important case where the auxiliary data in Table 1 include information about mutually exclusive and exhaustive subsets of the population U , and of the first-phase sample s_1 . The population subsets are denoted by U_i , $i = 1, \dots, I$, and the first-phase subsets by s_{1j} , $j = 1, \dots, J$. Such subsets are called calibration groups, for reasons that will become clear later in this Section. Simple examples of calibration groups are poststrata.

Two vectors denoted Δ_{1k} and Δ_{2k} will be used to specify the membership of a given unit k in the calibration groups U_i and s_{1j} , respectively. These group identifiers are

$$\Delta_{1k} = (\delta_{11k}, \dots, \delta_{1Ik}, \dots, \delta_{1Jk})' \quad (5.1)$$

with

$$\delta_{1ik} = \begin{cases} 1 & \text{if } k \in U_i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \dots, I \quad (5.2)$$

and

$$\Delta_{2k} = (\delta_{21k}, \dots, \delta_{2jk}, \dots, \delta_{2Ik})' \quad (5.3)$$

with

$$\delta_{2jk} = \begin{cases} 1 & \text{if } k \in s_{1j} \\ 0 & \text{otherwise} \end{cases} \quad \text{for } j = 1, \dots, J \quad (5.4)$$

Besides the group membership information, which is qualitative and specified by Δ_{1k} and Δ_{2k} , there may exist information for the unit k about quantitative (continuous or discrete) variables. We call them *supplementary auxiliary variables*. For example, categorical information about a unit (enterprise) in a business survey may consist of an industry code or a geographical location code. In addition, quantitative variable information may also be available concerning the number of employees or the gross business income of the unit. Some of these supplementary auxiliary variables may be known up to the level of the population, and others up to the level of the first-phase sample.

We assume in this Section that the vector \mathbf{x}_{1k} , used in calculating the first-phase g -factors, has the structure

$$\mathbf{x}'_{1k} = \Delta'_{1k} \otimes \mathbf{z}'_{1k} \quad (5.5)$$

where \mathbf{z}_{1k} of dimension Q_1 is the vector of supplementary auxiliary variables available for the first-phase sample. The information requirements in Table 1 apply to the vector \mathbf{x}_{1k} . This implies that we must know either the group membership specified by Δ_{1k} and the value of \mathbf{z}_{1k} for every $k \in U$, or the total $\sum_{U_i} \mathbf{z}_{1k}$ separately for each group, $i = 1, \dots, I$.

When \mathbf{x}_{1k} has the form given by (5.5), the first-phase g -factors g_{1k} in (3.5) can be obtained by a group by group calculation. The \mathbf{T}_1 matrix to be inverted, given by (3.6), is block diagonal and of dimension $I Q_1$ by $I Q_1$. The typical diagonal block, denoted as \mathbf{T}_{1i} of dimension Q_1 by Q_1 , is given by

$$\mathbf{T}_{1i} = \sum_{s_{1i}} \frac{w_{1k} \mathbf{z}_{1k} \mathbf{z}'_{1k}}{C_{1k}} \quad (5.6)$$

for $i = 1, \dots, I$. The resulting inverse of \mathbf{T}_1 is also block diagonal with diagonal matrices \mathbf{T}_{1i}^{-1} . The off diagonal blocks of the inverse of \mathbf{T}_1 are zero matrices. So we obtain from (3.6)

$$g_{1k} = 1 + \left(\sum_{U_i} \mathbf{z}_{1k} - \sum_{s_{1i}} w_{1k} \mathbf{z}_{1k} \right)' \mathbf{T}_{1i}^{-1} \frac{\mathbf{z}_{1k}}{C_{1k}} \quad (5.7)$$

for $k \in s_{1i}$, $i = 1, \dots, I$, where \mathbf{T}_{1i} is given by (5.6). Note that the resulting weights \tilde{w}_{1k} are the same as those obtained by carrying out the first-phase calibration group by group, calibrating for group i on the known total $\sum_{U_i} \mathbf{z}_{1k}$. That is, $\sum_{s_{1i}} \tilde{w}_{1k} \mathbf{z}_{1k} = \sum_{U_i} \mathbf{z}_{1k}$ for $i = 1, \dots, I$. It is thus fitting to call the groups U_i *first-phase calibration groups*.

Now consider the second-phase g -factors g_{2k} given by (3.11). They are based on the auxiliary vectors \mathbf{x}_k , required to be known for the units $k \in s_1$. We assume that \mathbf{x}_k contains information about the second-phase groups so that

$$\mathbf{x}'_k = \Delta'_{2k} \otimes \mathbf{z}'_k \quad (5.8)$$

where Δ_{2k} is the second-phase group identifier, and \mathbf{z}_k is the value of a vector of supplementary auxiliary variables available for $k \in s_1$. Since the requirements in Table 1 apply, it follows that Δ_{2k} (the second-phase group membership) and the value of \mathbf{z}_k (the supplementary auxiliary vector) must be known for every $k \in s_1$. Here \mathbf{z}_k may contain some or all of the information in \mathbf{x}_{1k} given by (5.5), and any other information available for the units $k \in s_1$.

When \mathbf{x}_k has the structure (5.8), the factors g_{2k} can also be obtained through a group by group calculation. This simplification is a result of the fact that the matrix to be inverted in (3.11) is block diagonal. We obtain

$$g_{2k} = 1 + \left(\sum_{s_{1j}} \tilde{w}_{1k} \mathbf{z}_k - \sum_{s_{2j}} \tilde{w}_{1k} w_{2k} \mathbf{z}_k \right)' \mathbf{T}_{2j}^{-1} \frac{\mathbf{z}_k}{C_{2k}} \quad (5.9)$$

for $k \in s_{2j} = s_2 \cap s_{1j}$, $j = 1, \dots, J$, where

$$\mathbf{T}_{2j} = \sum_{s_{2j}} \frac{\tilde{w}_{1k} w_{2k} \mathbf{z}_k \mathbf{z}'_k}{C_{2k}} \quad (5.10)$$

The resulting overall weights $\tilde{w}_k^* = w_k^* g_k^*$ where $g_k^* = g_{1k} g_{2k}$ are the same as those obtained by carrying out the second-phase calibration group by group, calibrating for group j on the known quantity $\sum_{s_{1j}} \tilde{w}_{1k} \mathbf{z}_k$. That is, $\sum_{s_{2j}} \tilde{w}_k^* \mathbf{z}_k = \sum_{s_{1j}} \tilde{w}_{1k} \mathbf{z}_k$ for $j = 1, \dots, J$. The groups s_{1j} are called *second-phase calibration groups*. We now have a procedure for computing g_{1k} and g_{2k} group by group using (5.7) and (5.9). The total Y is still estimated according to (3.13).

6. DOMAIN ESTIMATION AND VARIANCE ESTIMATION

The preceding sections dealt with estimation of the total of y at the entire population level. In most surveys, there is also a need to provide estimates for various subpopulations or domains of interest. Requests for domain estimates can be made either before or after the sampling stage of the survey. Auxiliary information is essential for domains. A

precise domain estimate may be obtained (even for small domains) if: (i) calibration groups and domains of interest agree closely, and (ii) the auxiliary variables exhibit a strong regression relationship with the variable(s) of interest.

Denote by $U_d (U_d \subseteq U)$ any domain of the population U for which an estimate is required. The y -total for the domain U_d is defined by $Y(d) = \sum_{U_d} y_k = \sum_U y_k(d)$ with $y_k(d) = y_k$ if $k \in U_d$ and $y_k(d) = 0$ if $k \notin U_d$.

The estimator of $Y(d)$ is

$$\hat{Y}(d) = \sum_{s_2} \tilde{w}_k^* y_k(d) \quad (6.1)$$

where the overall calibrated weights $\tilde{w}_k^* = w_k^* g_k^*$ may be calculated group by group as described in Section 5. The calibration factors g_{1k} and g_{2k} are calculated using all relevant available auxiliary information, specified as in Table 1. So in this sense, the resulting overall calibrated weights \tilde{w}_k^* are the best possible ones. Note that these weights are independent of the particular domains requiring estimation in the survey.

The estimator of the variance for the domain total estimator $\hat{Y}(d)$ is obtained using a design-based approach. This means that the variance is interpreted with reference to repeated draws of samples s_1 and s_2 . Details for the derivation of this variance are given in Särndal *et al.* (1992) (Result 9.7.1, p. 362). The first order and second order inclusion probabilities enter into the weights used in the variance formula. The weights associated with the first-phase sample are $w_{1k} = 1/\pi_{1k}$ and $w_{1kl} = 1/\pi_{1kl}$ with $\pi_{1kl} = P(k \text{ and } l \in s_1)$. The weights $w_{2k} = 1/\pi_{2k}$ and $w_{2kl} = 1/\pi_{2kl}$ with $\pi_{2kl} = P(k \text{ and } l \in s_2 | s_1)$ denote their second phase counterparts. Two sets of regression residuals, one for each phase, are also required. The estimator of the variance of $\hat{Y}(d)$ is given by

$$\begin{aligned} v\{\hat{Y}(d)\} = & \sum_{k \in s_2} \sum_{\ell \in s_2} w_{2kl} (w_{1k} w_{1\ell} - w_{1kl}) (g_{1k} e_{1k}(d)) (g_{1\ell} e_{1\ell}(d)) + \\ & \sum_{k \in s_2} \sum_{\ell \in s_2} w_{1k} w_{1\ell} (w_{2k} w_{2\ell} - w_{2kl}) (g_{2k} e_{2k}(d)) (g_{2\ell} e_{2\ell}(d)) \end{aligned} \quad (6.2)$$

Note that for $k = \ell$ we have $w_{1kl} = w_{1k}$, and $w_{2kl} = w_{2k}$ in (6.2). We now specify the regression residuals in (6.2) assuming that there are first-phase calibration groups $U_i, i = 1, \dots, I$, and second-phase calibration groups $s_{1j}, j = 1, \dots, J$, as explained in Section 5. We denote the associated sample subsets as follows: $s_{2i} = s_2 \cap U_i$; $s_{2j} = s_2 \cap s_{1j}$. The required residuals in (6.2) are, for $k \in (s_{2i} \cap U_d)$,

$$e_{1k}(d) = y_k(d) - \mathbf{z}'_{1k} \hat{\mathbf{B}}_{1i}(d) \quad (6.3)$$

and, for $k \in (s_{2j} \cap U_d)$

$$e_{2k}(d) = y_k(d) - \mathbf{z}'_k \hat{\mathbf{B}}_{2j}(d) \quad (6.4)$$

The estimated regression vectors $\hat{\mathbf{B}}_{1i}(d)$ and $\hat{\mathbf{B}}_{2j}(d)$ are

$$\hat{\mathbf{B}}_{1i}(d) = \mathbf{T}_{1i}^{-1} \left\{ \sum_{s_{1i}} \frac{w_{1k} \mathbf{z}_{1k} \hat{y}_{2k}(d)}{C_{1k}} + \sum_{s_{2i}} \frac{w_k^* \mathbf{z}_{1k} (y_k(d) - \hat{y}_{2k}(d))}{C_{1k}} \right\} \quad (6.5)$$

where \mathbf{T}_{1i} is given by (5.6), and

$$\hat{\mathbf{B}}_{2j}(d) = \mathbf{T}_{2j}^{-1} \sum_{s_{2j}} \frac{\tilde{w}_{1k} w_{2k} \mathbf{z}_k y_k(d)}{C_{2k}} \quad (6.6)$$

with \mathbf{T}_{2j} given by (5.10), and

$$\hat{y}_{2k}(d) = \mathbf{z}'_k \hat{\mathbf{B}}_{2j}(d) \text{ for } k \in (s_{1j} \cap U_d).$$

Remark 6.1: Note that for each new domain of interest, the variance estimator (6.2) requires two new sets of domain dependent residuals, $e_{1k}(d)$ and $e_{2k}(d)$. Moreover, these are required for *all* of the units k in the second-phase sample s_2 , including units outside the domain. Variance estimation for domains can therefore be cumbersome.

Remark 6.2: In practice the computation of estimated variances is seldom carried out as a double sum. For some important designs, the double sums reduce, after some algebraic manipulation, to single sum expressions. Examples of this occur for single sampling and for stratified single random sampling in both phases. Explicit algebraic developments for the variances have been given the former case by Särndal *et al.* (1992), and in the later case by Hidirolou (1995), and Binder, Babyak, Brodeur, Hidirolou and Jocelyn (1997).

7. APPLICATIONS WITH POSTSTRATIFICATION AT THE FIRST PHASE

7.1 The Case of the Tax Sample at Statistics Canada

An application of the calibration group approach in section 5 has been in use at Statistics Canada, in the two-phase design for sampling of tax records. The example is important because it provides the extension to two-phase designs of the traditional poststratification technique as used in a single phase design. The sampling procedure, the post-stratification criteria, and the estimators are described in Armstrong and St-Jean (1994). We now show how these estimators are obtained as special case of the technique in section 5. The sampling design, in each phase, is stratified Bernoulli, carried out with the permanent random number technique. The two stratifications are based on different criteria. The realized sample sizes are random at each phase on account of the Bernoulli sampling. To offset the resulting tendency toward an increased variance, poststratification is carried out at both phases of sampling. The two

poststratification criteria are different. We have in effect two crossing poststratifications. In the terminology of section 5, the first phase poststrata are the first-phase calibration groups. They are denoted as U_i ; $i = 1, \dots, I$, and the group membership of a unit k is indicated by the vector by Δ_{1k} given by (5.1). The second phase poststrata are the second phase calibration groups. They are denoted as s_{1j} , $j = 1, \dots, J$ and the corresponding membership of a unit k is indicated by the vector Δ_{2k} given by (5.3).

The first-phase calibration is carried out using the information about the first-phase poststrata sizes, N_i . In this survey design, there is no supplementary information, so $z_{1k} = 1$ for all k in (5.5), yielding $x_{1k} = \Delta_{1k}$. Specifying $C_{1k} = 1$ for all k we obtain from (5.7) that

$$g_{1k} = N_i / \hat{N}_{1i} \quad (7.1)$$

for all $k \in s_{1i}$ where $\hat{N}_{1i} = \sum_{k \in s_{1i}} w_{1k}$ estimates the known first-phase poststratum count N_i , and $s_{1i} = s_1 \cap U_i$ denotes the part of the first-phase sample s_1 that falls in the first-phase poststratum U_i .

We arrive at the estimator of Armstrong and St-Jean (1994) by carrying out the second-phase calibration with $x_k = \Delta_{2k}$, that is, we have $z_k = 1$ for all k in (5.8). This is a reduced x_k -vector specification since it does not involve x_{1k} . Specifying $C_{2k} = 1$ for all $k \in s_{1i}$, and using (5.9) and (3.10), we obtain the overall calibrated weights

$$g_k^* = \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1ij}}{\hat{N}_{2ij}} \quad (7.2)$$

for all $k \in s_{2ij}$, where

$$\hat{N}_{1ij} = \sum_{i=1}^I \left(\frac{N_i}{\hat{N}_{1i}} \right) \hat{N}_{1ij}; \hat{N}_{2ij} = \sum_{i=1}^I \left(\frac{N_i}{\hat{N}_{1i}} \right) \hat{N}_{2ij} \quad (7.3)$$

with $\hat{N}_{1ij} = \sum_{k \in s_{1ij}} w_{1k}$ and $\hat{N}_{2ij} = \sum_{k \in s_{2ij}} w_k^*$. Here, $s_{2ij} = s_2 \cap s_{1ij}$ denotes the part of the second-phase sample s_2 that falls in the second-phase poststratum s_{1j} , and $s_{1ij} = U_i \cap s_{1j}$; $s_{2ij} = s_2 \cap U_i \cap s_{1j}$. It follows that the estimator of the total $Y(d)$ for a given domain U_d is given by $\hat{Y}(d) = \sum_{s_2} w_k^* g_k^* y_k(d)$, or equivalently as

$$\hat{Y}(d) = \sum_{i=1}^I \sum_{j=1}^J \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1ij}}{\hat{N}_{2ij}} \sum_{s_{2ij}} w_k^* y_k(d).$$

The estimated variance requires two types of residuals that are easily obtained from the general expressions given in Section 6.

Alternatives exist to the reduced vector specification $x_k = \Delta_{2k}$ used for this design. We therefore examine what the estimator would look like under a full vector specification. For the first-phase calibration, as earlier, let $x_{1k} = \Delta_{1k}$ corresponding to $z_{1k} = 1$ for all k in (5.8). The first-phase g -factors g_{1k} are then given by (7.1). In this

survey, information is available for assigning every unit $k \in s_1$ to one of the $I \times J$ cells formed by cross-classifying the two poststratification criteria. Therefore, the vector x_k for the second-phase calibration can be taken as

$$x_k' = \Delta_{1k} \otimes \Delta_{2k}' \quad (7.4)$$

This is a full vector specification in that it includes the first-phase information carrier Δ_{1k} . Let us also specify $C_{2k} = 1$ for all k . Since (7.4) is of the form (5.8), the second-phase g -factors g_{2k} are obtainable group-by-group from (5.9) with $z_k = \Delta_{1k}$. The overall calibration factors are given by

$$g_k^* = \frac{N_i}{\hat{N}_{1i}} \frac{\hat{N}_{1ij}}{\hat{N}_{2ij}} \quad (7.5)$$

for all $k \in s_{2ij}$. Here, \hat{N}_{1i} is defined in (7.1), and \hat{N}_{1ij} and \hat{N}_{2ij} are as in (7.3). These overall calibration factors are the product of two poststratified calibration factors. They are all positive and well defined, provided all sample cells s_{2ij} are non-empty. Collapsing of small cells s_{2ij} with relatively large non-empty cells is recommended for stable estimation. As pointed out in Remark 3.4, the overall weights obtained from (7.5) reproduce the known first-phase poststrata sizes N_i , whereas those obtained from (7.2) do not.

Remark 7.1: Let us compare the calibration factors (7.2) and (7.5), resulting, respectively, from the reduced form $x_k = \Delta_{2k}$ and from the full form (7.4). Both factors are a product of two terms. The only difference lies in the second term. In both cases, the computation of the second term requires cross-classification information. That is, for every $k \in s_1$, we need to identify the cross-classification cell ij to which k belongs. In the case of the reduced vector, the cell information is pooled across the first-phase groups. For the full vector, the cell information is kept separate, and one would expect the resulting weights to be more efficient.

Remark 7.2: For the second-phase calibration, an alternative to (7.4) that also captures the information about the first-phase poststrata is to use

$$x_k' = (\Delta_{1k}', \Delta_{2k}'). \quad (7.6)$$

Note that with this specification, there is only one calibration group in the second phase, namely the whole first-phase sample s_1 .

7.2 The Case of the Canadian Survey Employment, Payrolls and Hours

The Survey on Employment Payrolls, and Hours (SEPH) covers all sectors of Canadian industry, and collects data on four principal variables: (i) salaries and payments to employees (denoted as z_2 ; called payrolls); (ii) number of employees (z_3 ; employment); (iii) hours worked by employees (y_1 ; hours); and (iv) summarized earnings (y_2 ; earnings).

SEPH (1994) uses a stratified two-phase sampling design. In the first phase, a sample of payroll deduction accounts is selected using a stratified Bernoulli sampling design with sampling rates within strata ranging from 10% to 100%. The strata are defined by region. A region is made up of one or more Canadian provinces. We describe the estimation for SEPH by considering one specific region.

For units selected in the first-phase sample, two variables are transcribed, namely, payrolls (z_2) and number of employees (z_3). In the second-phase, a simple random sample is drawn. Data on the two variables of interest, y_1 and y_2 , are collected for respondents in this sample. In addition, classification by industry and province is recorded for sampled units. The first-phase sample is poststratified by employment size groups. These are used as first-phase calibration groups and denoted U_i ; $i = 1, \dots, I$. Their sizes denoted as N_i for $i = 1, \dots, I$ are assumed known. The vector \mathbf{x}_{1k} used for a first-phase calibration is of the form (5.5), where Δ_{1k} is given by (5.1) and $z_{1k} = 1$ for all k . We choose $C_{1k} = 1$ for all k . It follows from (5.7) that the first-phase g -factors are

$$g_{1k} = N_i / \hat{N}_{1i} \quad (7.7)$$

for all $k \in s_{1i} = s_i \cap U_i$, where $\hat{N}_{1i} = \sum_{s_{1i}} w_{1k}$, $i = 1, \dots, I$.

We now turn to second-phase calibration. It is carried out using calibration groups s_{1j} , $j = 1, \dots, J$, identified by the vector Δ_{2k} given by (5.3). These groups are based on a province by industry classification. They are constructed so that: (i) there is a strong regression relationship between y_k and the two z -variables, and that (ii) there are at least 30 observations within each group. The $J(I+2)$ dimensional \mathbf{x}_k -vector for the second-phase calibration is given by

$$\mathbf{x}'_k = \Delta'_{2k} \otimes (\Delta'_{1k}, z_{2k}, z_{3k}) \quad (7.8)$$

This specification requires (see Table 1) that every $k \in s_1$ can be classified into one of the I by J cells formed by crossing the calibration groups in the two phases. Let $s_{2j} = s_2 \cap s_{1j}$; $s_{1ij} = s_{1j} \cap U_i$; $s_{2ij} = s_{2j} \cap s_{1ij}$. Also, the quantitative variable values z_{2k} (payrolls) and z_{3k} (number of employees) must be known for $k \in s_1$. The \mathbf{x}_k -vector specification given by (7.8) is full, because it incorporates $\mathbf{x}_{1k} = \Delta_{1k}$. A reduced vector, ignoring the first-phase groups, would be $\mathbf{x}'_k = \Delta'_{2k} \otimes (z_{2k}, z_{3k})$.

As in Example 7.1, we have two crossing sets of calibration groups.

Since the \mathbf{x}_k -vector (7.8) has the structure defined by (5.8), we used (5.9) to derive the second-phase g -factors for each group $j = 1, \dots, J$. It follows from (7.8) that we are fitting, within each second-phase calibration group, a separate regression of y_k on $\zeta_k = (z_{2k}, z_{3k})'$ with an intercept that varies with the first-phase calibration group.

Specifying $C_{2k} = 1$ for all k , and using the additive form, $g_k^* = g_{1k} + g_{2k} - 1$, for the overall calibration factors, we obtain after some algebra

$$g_k^* = G_{1i} G_{2ij} + \mathbf{H}'_j \mathbf{T}_j^{-1} (\zeta_k - \bar{\zeta}_{s_{2ij}})$$

for all $k \in s_{2ij}$, where

$$G_{1i} = N_i / \hat{N}_{1i}, G_{2ij} = \hat{N}_{1ij} / \hat{N}_{2ij}$$

$$\mathbf{H}_j = \sum_{i=1}^I \hat{N}_{1ij} G_{1i} (\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}})$$

$$\mathbf{T}_j = \sum_{i=1}^I \sum_{s_{2ij}} w_k^* (\zeta_k - \bar{\zeta}_{s_{2ij}}) (\zeta_k - \bar{\zeta}_{s_{2ij}})'$$

with

$$\bar{\zeta}_{s_{1ij}} = \sum_{s_{1ij}} \frac{w_{1k} \zeta_k}{\hat{N}_{1ij}}; \bar{\zeta}_{s_{2ij}} = \sum_{s_{2ij}} \frac{w_k^* \zeta_k}{\hat{N}_{2ij}}; \hat{N}_{1ij} = \sum_{s_{1ij}} w_{1k};$$

$$\text{and } \hat{N}_{2ij} = \sum_{s_{2ij}} w_k^*.$$

It follows that we can write the estimator (6.1) as $\hat{Y}(d) = \sum_{i=1}^I \sum_{j=1}^J \hat{Y}_{ij}(d)$ with

$$\hat{Y}_{ij}(d) = G_{1i} \hat{N}_{1ij} \{ \bar{y}_{s_{2ij}}(d) + (\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}})' \hat{\mathbf{B}}_j(d) \}$$

where

$$\bar{y}_{s_{2ij}}(d) = \sum_{s_{2ij}} w_k^* y_k(d) / \hat{N}_{2ij}$$

$$\text{and } \hat{\mathbf{B}}_j(d) = \mathbf{T}_j^{-1} \sum_{i=1}^I \sum_{s_{2ij}} w_k^* (\zeta_k - \bar{\zeta}_{s_{2ij}}) y_k(d).$$

The form of $\hat{Y}(d)$ is easy to understand. It is composed of $I \times J$ cell estimates $\hat{Y}_{ij}(d)$, each reflecting the regression of $y_k(d)$ on ζ_k . Note that the two-dimensional slope vector $\hat{\mathbf{B}}_j(d)$ is obtained by pooling data across the first-phase groups. This is because the specification (7.8) of \mathbf{x}_k allows the intercept, but not the two regression slopes, to vary with the first-phase groups.

8. CONCLUSIONS

Two-phase designs have the advantage of being both economical and efficient. The present paper has provided a general theory for such designs when auxiliary information is present in each phase.

Our goal is to incorporate this two-phase survey methodology into Statistics Canada's Generalized Estimation System (GES) described in Estevao *et al.* (1995). The GES is a general purpose program that currently handles domain estimation for arbitrary single phase designs and incorporates auxiliary information in its estimation process. In this paper we have extended the basic principles of the GES, including the important idea of calibration groups, to two-phase designs.

We have illustrated the theory by showing its use in two current surveys at Statistics Canada. Given its generality, the theory has potential application to any two-phase sample design that uses auxiliary information.

REFERENCES

- ARMSTRONG, J., and ST-JEAN, H. (1994). Generalized regression estimation for a two-phase sample of tax records. *Survey Methodology*, 20, 97-106.
- BINDER, D.A. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22, 17-22.
- BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A., and JOCELYN, W. (1997). Variance Estimation for Two-phase Stratified Sampling. Contributed paper presented at the Annual Meeting of the American Statistical Association, Los Angeles.
- BREIDT, J., and FULLER, W.A. (1993). Regression weighting for multiphase samples. *Sankhyā*, 55, 297-309.
- CHAUDHURI, A., and ROY, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: John Wiley.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAL, C.-E., and SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DUPONT, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, 21, 125-136.
- ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- HIDIROGLOU, M.A. (1995). Sampling and estimation for stage one of the Canadian Survey of Employment, Payrolls and Hours redesign. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 123-128.
- HIDIROGLOU, M.A., LATOUCHE, M., ARMSTRONG, B., and GOSSEN, M. (1995). Improving survey information using administrative records: the case of the Canadian employment survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.
- HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 873-878.
- NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of The American Statistical Association*, 33, 101-116.
- SÄRNDAL, C.-E., and SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New-York: Springer-Verlag.
- SINGH, A.C., and MOHL, C.A. (1996). Understanding calibration estimation in survey sampling. *Survey Methodology*, 22, 107-115.
- STUKEL, D., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus linearization. *Survey Methodology*, 22, 117-125.

Estimation in Sample Surveys Using Frames With a Many-to-Many Structure

TERRI L. BYCZKOWSKI, MARTIN S. LEVY and DENNIS J. SWEENEY¹

ABSTRACT

In sample surveys, the units contained in the sampling frame ideally have a one-to-one correspondence with the elements in the target population under study. In many cases, however, the frame has a many-to-many structure. That is, a unit in the frame may be associated with multiple target population elements and a target population element may be associated with multiple frame units. Such was the case in a building characteristics survey in which the frame was a list of street addresses, but the target population was commercial buildings. The frame was messy because a street address corresponded either to a single building, multiple buildings, or part of a building. In this paper, we develop estimators and formulas for their variances in both simple and stratified random sampling designs when the frame has a many-to-many structure.

KEY WORDS: Imperfect frames; Correspondence errors; Building characteristics survey; Weighting; Simple random sampling; Stratified random sampling.

1. INTRODUCTION

This research was motivated by a study that was conducted for a utility company to estimate various population characteristics of the commercial buildings located in their service area. Budgetary constraints prohibited the development of a list of commercial buildings using canvassing techniques. However, a sampling frame consisting of street addresses (*i.e.*, addresses at which a utility meter was located) was available. A drawback of this frame was that it had a many-to-many relationship with the target population of commercial buildings. That is, some units in the frame were associated with multiple target population elements, and some target population elements were associated with multiple frame units. In fact, several of the relationships between street addresses and commercial buildings were relatively complex.

An advantage of this frame, however, was that total annual electrical usage was available for each street address. This resulted in a variable upon which the frame of street addresses could be effectively stratified. One of the important characteristics to be measured was the total commercial square footage. Studies conducted in the United States have shown that energy consumption is associated with both building size and building activity. For example, consumption is higher for buildings used for health care or food sales, and lower for buildings used for religious worship or public assembly. Also, energy consumption is correlated with building size even if the activity of the building is not known, as was the case here (U.S. Department of Energy 1992).

There is a vast amount of literature dealing with imperfect sampling frames. Comprehensive summaries of this literature can be found in Kish (1965), Wright and Tsao

(1983), and Lessler and Kalsbeek (1992). Another body of literature addresses multiplicity sampling in which the frame is constructed with a many-to-many structure by design. Here, frame imperfections are introduced in order to gather information more efficiently on rare occurrences in a population (Birnbaum and Sirken 1965, Sirken 1972a,b, and Casady and Sirken 1980). Hansen, Hurwitz and Madow (1953a,b) present an estimator for use with sampling frames that have a many-to-one structure; population elements are represented multiple times in the frame. This estimator has also been adopted for use by National Agricultural Statistics Service (NASS) surveys (Musser 1993) with respect to the many-to-one frame. Bandyopadhyay and Adhikari (1993) developed estimators for a ratio, population mean, and population total when an unknown amount of duplication is present in the frame. But, these estimators are restricted to the simple random sampling case and the many-to-one frame.

Two methods for estimating population characteristics using a frame with a many-to-many structure appear in the literature. First, the Horvitz-Thompson estimator (1952) provides unbiased estimates of population means and totals when varying probabilities of selection are present. Musser (1993) shows how to compute the correct inclusion probabilities for the population elements selected in simple random sampling from a many-to-one frame. However, Musser's method can be extended to obtain inclusion probabilities for population elements in a simple random sample from the many-to-many frame as well. Second, Lavallée (1995) adapted the Weight Share Method, applied to longitudinal surveys, to the use of frames with a many-to-many structure.

The purpose of this paper is to develop an alternative methodology for estimating population totals, counts, and

¹ Terri L. Byczkowski, Institute for Policy Research, Martin S. Levy and Dennis J. Sweeney, Department of Quantitative Analysis and Operations Management, University of Cincinnati, Cincinnati, OH 45221, U.S.A.

means when using sampling frames with a many-to-many structure under simple and stratified random sampling designs. Also, expressions for the variance of those estimators are derived. The results which we develop are not only of intrinsic interest, but expressions for the variance of the estimators are essential for the exploration of the effects of correspondence imperfections inherent in many-to-many sampling frames on the precision of these estimates.

In section 2 we present these estimates in the simple random sampling without replacement (SRSWOR) case. We also describe the sampling methodology under which these estimators are applicable, state a result on bias, and develop expressions for their variance.

In section 3 some of the results are extended to the case of stratified random sampling. In section 4 we develop conclusions, discuss limitations and make suggestions for future research.

2. MANY-TO-MANY FRAMES FOR SIMPLE RANDOM SAMPLING

It is useful to think of the relationship between the frame and the target population as a graph. The sampling units in the frame and the elements of the target population are the two sets of nodes; arcs link the sampling units to elements of the target population. These arcs reveal the structure of the relationship between the frame and the target population. Figure 2.1 shows an example of a frame and target population with a many-to-many relationship. There are 7 sampling units in the frame, 6 elements in the target population and 10 links (arcs) between the sampling units and the elements of the population. Thus, a graph with 13 nodes and 10 arcs represents this many-to-many structure. In this paper we assume that each population element is linked to the set of frame units by at least one arc and that each frame unit is linked to the set of population elements by at least one arc as well.

Let us fix some notation. We find it convenient to identify both frame units and population elements with their respective indices. Let $F = \{1, 2, ..., N\}$ denote the set of indices for N sampling units, and let $T = \{1, 2, ..., M\}$ denote the set of indices for the M target population elements. An arc can be represented as an ordered pair; the first element of which comes from F , and the second from T . A population element k in T is said to be represented by sampling unit j in F , if it is linked to it by an arc denoted (jk) . This means that when j is in the sample there is a nonzero probability of collecting data from population element k . We will denote by y_k the measurement of interest on target population element k in T .

We now describe the sampling methodology under which the estimators developed herein are appropriate. Assume a SRSWOR of size n frame units is selected from F . The number of *population elements* included in the sample and measured, however, depends upon the nature of

the association between the frame units and the population elements.

Under SRSWOR, one of four scenarios can occur when a frame unit is selected. In the first scenario, a frame unit corresponds to one and only one population element (a one-to-one structure). Here the surveyor would simply collect the information concerning the single population element corresponding to the selected frame unit (see frame unit 1 of Figure 2.1).

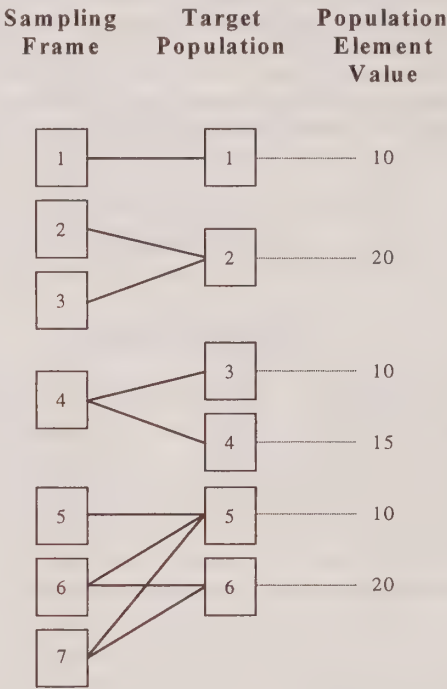


Figure 2.1. An example of the correspondence between the sampling frame and the target population

In the second scenario, several frame units correspond to one population element (a many-to-one structure). For example, in Figure 2.1, frame units 2 and 3 correspond to the single population element 2. In this case, if frame units 2 and/or 3 are included in the sample, information on population element 2 is collected. Thus, it is possible that population element 2 could appear in the sample, and as a record in the data set used to develop the estimates, up to two times.

In the third scenario, one frame unit corresponds to more than one population element (a one-to-many structure). For example, in Figure 2.1 frame unit 4 corresponds to population elements 3 and 4. Here, only one population element (3 or 4) is selected using a *randomization* independent of the choice of frame units. Economics dictated this policy because data collection entailed lengthy personal interviews conducted by individuals with technical backgrounds. In this paper we assume that these randomizations are conducted using equal probabilities. But, any probabilities could be used (e.g., probability proportional to size) provided they are non-zero.

In the fourth scenario, a many-to-many structure exists. This is illustrated by frame units 5, 6 and 7 and population elements 5 and 6 in Figure 2.1. Since these complex cases are combinations of scenarios 2 and 3 above, the same sampling rules apply. For example, if frame unit 5 is selected, population element 5 is measured. If frame unit 6 is selected, only one of population elements 5 and 6 is randomly selected and measured.

2.1 Population Totals

2.1.1 Estimator for a Population Total

A many-to-many frame results in varying probabilities of selection. The estimators developed here involve a method of weighting, which is an extension of the estimator presented by Hansen *et al.* (1953a pp. 62-64). Their estimators and formulas for the variance of those estimators are restricted to the many-to-one frame structure. We extend those estimators to the many-to-many frame structure.

For a SRSWOR of size n , let J_1, \dots, J_n denote random variables such that $J_i = j$ if the i -th draw results in the selection of unit j from F . Hence $\Pr(J_i = j) = 1/N$ for j in F and $i = 1, \dots, n$. Let K_1, \dots, K_n denote random variables such that $K_i = k$ if the i -th draw from F is followed by the selection of k from T . We can now think of drawing a random sample of arcs $\{(J_1 K_1), \dots, (J_n K_n)\}$ which has a joint probability distribution determined by both the SRSWOR sampling design and the subsequent randomization (if required) to choose an element in T . In particular, $(J_i K_i)$ has marginal probability given by $\Pr\{(J_i K_i) = (jk)\} = (1/N)s_{jk}$, in which s_{jk} is the conditional probability given by, $s_{jk} = \Pr(K_i = k | J_i = j)$. That is, s_{jk} is the conditional probability of selecting population element k in T given that frame unit j in F is selected. These conditional probabilities will be referred to as arc probabilities and are illustrated for Figure 2.1 in Table 2.1.

Table 2.1
Arc Probabilities for Figure 2.1

Arc jk	1,1	2,2	3,2	4,3	4,4	5,5	6,5	6,6	7,5	7,6
s_{jk}	1	1	1	1/2	1/2	1	1/2	1/2	1/2	1/2

For k in T , let U_k denote the set of units in F that have arcs with a destination at k in T . Let $s_k = \sum_{j \in U_k} s_{jk}$. Using the language in Hansen *et al.* (1953a pp. 62-64) which motivated our development, we call s_k the *weight* for population element k in T . These weights for Figure 2.1 appear in Table 2.2.

Table 2.2
Calculation of the Population Element Weights (s_k) for Figure 2.1

k	1	2	3	4	5	6
(s_k)	1	2	1/2	1/2	2	1

Arc probabilities and weights are used to compute the marginal probabilities of the K_i , namely, $\Pr(K_i = k) =$

$\sum_{j \in U_k} (1/N)s_{jk} = (1/N)s_k$, where k is in T , and $i = 1, \dots, n$. Clearly, computing the arc probabilities is the key step in developing the correct weights for the data collected. It depends on properly ascertaining the graph structure for each sampling unit selected: a maximally connected (MC) subgraph. A connected subgraph is a subset of the nodes which are connected by a sequence of arcs. Maximal means that no node outside the subset is connected to a node belonging to the subset. There are 4 MC subgraphs in Figure 2.1. Each represents a different frame – population structure, namely, one-to-one, many-to-one, one-to-many, and many-to-many structure.

To develop the estimators it is not necessary to know the structure for the entire graph. It is only necessary to know the structure of the MC subgraphs to which *sampled* frame units belong.

We make the following observations about s_k and s_{jk} : (i) $s_k = W$ indicates that population element k has W times the probability of being selected on the i -th draw as that of a population element with a weight of one; (ii) $0 < s_k \leq N$, $k = 1, \dots, M$; (iii) $0 < s_{jk} \leq 1$, $j \in U_k$ and $k = 1, \dots, M$; (iv) with respect to the one-to-many frame structure, $s_{jk} = s_k$; (v) with respect to the many-to-one frame structure, $s_{jk} = 1$ for all k ; and (vi) $\sum_{k=1}^M \sum_{j=1}^N s_{jk} = N$.

Now, let x_1, \dots, x_M denote the weighted values associated with the indices in T . That is, let $x_k = y_k/s_k$. Define random variables x_{K_1}, \dots, x_{K_n} , associated with draws 1 through n from F , respectively, so that x_{K_i} takes the value x_k if $K_i = k$. Notice that we can write,

$$E(x_{K_i}) = \sum_{k=1}^M x_k \Pr(K_i = k) = \frac{1}{N} \sum_{k=1}^M \frac{y_k}{s_k} s_k = \frac{Y}{N}, \quad (2.1)$$

where $Y = \sum_{k=1}^M y_k$ is the true population total. We take as our estimator of the population total based upon a SRSWOR from a sampling frame with many-to-many structure,

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^n x_{K_i}. \quad (2.2)$$

Using (2.1) it follows that,

$$E(\hat{Y}) = E\left(\frac{N}{n} \sum_{i=1}^n x_{K_i}\right) = \frac{N}{n} \sum_{i=1}^n E(x_{K_i}) = \frac{N}{n} n \frac{Y}{N} = Y.$$

We thus obtain,

Theorem 2-1: The estimator (2.2) for a population total used in SRSWOR is unbiased.

Using Figure 2.1, we now give a simple example of the use of this estimator. Suppose a simple random sample of four frame units was selected from the frame depicted in Figure 2.1 (2, 3, 4, and 7) which ultimately resulted in the selection of population elements 2, 4, and 5. The estimator of the population total,

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^4 x_{K_i}, \text{ has value } \frac{7}{4} \left[\frac{20}{2} + \frac{20}{2} + \frac{15}{(1/2)} + \frac{10}{2} \right] = \frac{385}{4}.$$

The above estimator can also be used for a population count. We could estimate the size of the target population by letting $y_k = 1$ for all k . In addition, we could estimate the number of population elements that possess some characteristic by letting $y_k = 1$ for those population elements with the characteristic of interest and $y_k = 0$ for those without the characteristic.

2.1.2 Variance of the Estimator for a Population Total

First, some additional terminology and notation used in this section must be defined. Let P represent the set of all unordered pairs of arcs. We shall define an unordered pair of arcs as *inadmissible* if they cannot both be included in a sample. Formally let $Q = \{j \text{ in } F: \text{more than one arc emerges from } j\}$. Then $R' = \{[jk, j'k']: j \in Q \text{ and } k \neq k'\}$ is the set of unordered *inadmissible* pairs of arcs. Also, the set of unordered *admissible* pairs of arcs is the complementary set $R^* = P \setminus R'$.

To illustrate, consider Figure 2.1. The sampling methodology we employ requires that if frame unit 4 is selected, only one of population elements 3 and 4 can be included in the sample. Thus, $\{[4,3][4,4]\}$ is an unordered inadmissible pair of arcs. The other unordered inadmissible pairs of arcs in Figure 2.1 are $\{[6,5][6,6]\}$ and $\{[7,5][7,6]\}$. Thus, $R' = \{[4,3][4,4], [6,5][6,6], [7,5][7,6]\}$.

Theorem 2-2: The variance of the estimator (2.2) is,

$$V(\hat{Y}) = \frac{N}{n} \left[\sum_{k=1}^M \frac{y_k^2}{s_k} + 2 \frac{(n-1)}{(N-1)} \sum_{[jk, j'k'] \in R^*} \left(\frac{y_k s_{jk} y_{k'} s_{j'k'}}{s_k s_{k'}} \right) \right] - Y^2, \quad (2.3)$$

where the double sum is over all unordered *admissible* pairs of arcs $[jk, j'k']$.

Proof:

$$\begin{aligned} V(\hat{Y}) &= E \left[\left(\frac{N}{n} \sum_{i=1}^n x_{K_i} \right)^2 \right] - Y^2 \\ &= \frac{N^2}{n^2} E \left[\left(\sum_{i=1}^n x_{K_i} \right)^2 \right] - Y^2. \end{aligned} \quad (2.4)$$

Now,

$$E \left[\left(\sum_{i=1}^n x_{K_i} \right)^2 \right] = \sum_{i=1}^n E(x_{K_i}^2) + 2E \left(\sum_{i < i'} (x_{K_i} x_{K_{i'}}) \right). \quad (2.5)$$

One can write

$$E(x_{K_i}^2) = \sum_{k=1}^M [x_k^2 \Pr(K_i = k)] = \sum_{k=1}^M \frac{y_k^2}{s_k^2} \frac{s_k}{N} = \frac{1}{N} \sum_{k=1}^M \frac{y_k^2}{s_k}. \quad (2.6)$$

As mentioned in Section 2.1, we can think of selecting a sample of arcs which ultimately leads to the selection of population elements. Each arc (jk) is associated with a value $x_k = y_k/s_k$ of the population element k at its destination. Thus, we can rewrite the double summation in (2.5) as a summation over admissible unordered pairs of arcs, R^* .

$$\begin{aligned} 2E \left(\sum_{i'} \sum_{i < i'} (x_{K_i} x_{K_{i'}}) \right) &= \\ 2 \binom{n}{2} \sum_{[jk, j'k'] \in R^*} [(x_k x_{k'}) \Pr(K_i = k, K_{i'} = k')]. \end{aligned} \quad (2.7)$$

Now, by virtue of the independence of the randomization and the choice of frame units:

$$\Pr(\text{select } [jk, j'k'] \text{ in } R^*) = \Pr(\text{select } \{j, j'\} \text{ in } F)$$

$$\Pr(\text{select } [jk, j'k'] \text{ in } R^* \mid \text{select } \{j, j'\} \text{ in } F) = \frac{1}{\binom{N}{2}} s_{jk} s_{j'k'}.$$

Substituting into (2.7) results in,

$$\begin{aligned} n(n-1) \sum_{[jk, j'k'] \in R^*} \left[(x_k x_{k'}) \frac{1}{\binom{N}{2}} s_{jk} s_{j'k'} \right] &= \\ \frac{2n(n-1)}{N(N-1)} \sum_{[jk, j'k'] \in R^*} \left[\left(\frac{y_k s_{jk} y_{k'} s_{j'k'}}{s_k s_{k'}} \right) \right]. \end{aligned} \quad (2.8)$$

Now substituting (2.6) and (2.8) into (2.5) yields,

$$\begin{aligned} E \left[\left(\sum_{i=1}^n x_{K_i} \right)^2 \right] &= \frac{n}{N} \sum_{k=1}^M \frac{y_k^2}{s_k} + \\ \frac{2(n-1)}{(N-1)} \sum_{[jk, j'k'] \in R^*} \left(\frac{y_k s_{jk} y_{k'} s_{j'k'}}{s_k s_{k'}} \right). \end{aligned} \quad (2.9)$$

Finally substituting (2.9) into (2.4) gives the result (2.3).

Equation (2.3) is a generalization of the formula developed by Bandyopadhyay and Adhikari (1993) for the variance of the estimate of a population total in the case of the many-to-many frame structure. It can be shown that (2.3) reduces to their formula when the sampling frame is restricted to a many-to-one structure.

Corollary 2-1: An alternative form of the variance formula in Theorem 2-2 is:

$$V(\hat{Y}) = \frac{N}{n} \left[\sum_{k=1}^M \frac{y_k^2}{s_k} + \frac{(n-1)}{(N-1)} \left(\left(\sum_{jk} \frac{y_k}{s_k} s_{jk} \right)^2 - \sum_{jk} \left(\frac{y_k}{s_k} s_{jk} \right)^2 - 2 \sum_{[jk, j'k'] \in R'} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right) \right] - Y^2.$$

Proof:
Write,

$$\left(\sum_{jk} \frac{y_k s_{jk}}{s_k} \right)^2 = \sum_{jk} \left(\frac{y_k s_{jk}}{s_k} \right)^2 + 2 \sum_{[jk, j'k'] \in R^*} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} + 2 \sum_{[jk, j'k'] \in R'} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}}.$$

It follows that:

$$\sum_{[jk, j'k'] \in R^*} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} = \frac{1}{2} \left(\sum_{jk} \frac{y_k s_{jk}}{s_k} \right)^2 - \frac{1}{2} \sum_{jk} \left(\frac{y_k s_{jk}}{s_k} \right)^2 - \sum_{[jk, j'k'] \in R'} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}}.$$

Substituting the above expression into (2.3) provides the result.

This formula is computationally simpler. Note that (2.3) requires that the term

$$\left(\frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right)$$

be summed over all unordered admissible pairs of arcs (R^*), whereas this alternative formula only requires a summation over pairs of arcs that are inadmissible (R'). In most practical scenarios the number of admissible pairs of arcs will be far greater than the number of inadmissible pairs of arcs.

2.2 Population Means

2.2.1 Estimator for a Population Mean

The estimator for a population mean presented here extends the estimator presented by Hansen *et al.* (1953a) to the many-to-many frame structure.

Associated with the n draws from F , define random variables s_{K_i} and $z_{K_i} = 1/s_{K_i}$, so that s_{K_i} takes value s_k if

$K_i = k$ for $i = 1, \dots, n$ and $k = 1, \dots, M$. The estimator for a population mean,

$$\bar{Y} = \frac{1}{M} \sum_{k=1}^M y_k,$$

when using SRSWOR and a many-to-many frame is:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^n x_{K_i}}{\sum_{i=1}^n z_{K_i}}. \quad (2.10)$$

2.2.2 Mean Square Error (MSE) of the Estimator for a Population Mean

The estimator for a population mean is biased because it is a ratio estimator. But, it is well known that this bias becomes negligible for large samples and the bias is of order $1/n$ (Cochran 1977, p. 160).

Our approximation of the MSE requires a summation over R^{**} , the set of all *ordered admissible* pairs of arcs. Thus, if $[jk, j'k'] \in R^*$, then both $[jk, j'k'] \in R^{**}$ and $[j'k', jk] \in R^{**}$.

To approximate the mean square error of the estimator (2.10), we use

$$\begin{aligned} \text{MSE}(\hat{\bar{Y}}) &\approx \frac{M^2}{nN \left(\sum_{k=1}^M \frac{1}{s_k} \right)^2} \\ &\left[\left(\sum_{k=1}^M \frac{y_k^2}{s_k} + \frac{2(n-1)}{(N-1)} \sum_{[jk, j'k'] \in R^*} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right) \right. \\ &- 2\bar{Y} \left(\sum_{k=1}^M \frac{y_k}{s_k} + \frac{(n-1)}{(N-1)} \sum_{[jk, j'k'] \in R^{**}} \frac{y_k s_{jk}}{s_k} \frac{y_{k'} s_{j'k'}}{s_{k'}} \right) \\ &\left. + \bar{Y}^2 \left(\sum_{k=1}^M \frac{1}{s_k} + \frac{2(n-1)}{(N-1)} \sum_{[jk, j'k'] \in R^*} \frac{s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}} \right) \right]. \quad (2.11) \end{aligned}$$

To justify this approximation let

$$\bar{x} = \frac{\sum_{i=1}^n x_{K_i}}{n}, \quad \bar{z} = \frac{\sum_{i=1}^n z_{K_i}}{n} \quad \text{and} \quad \bar{Z} = \frac{\sum_{k=1}^M \frac{1}{s_k}}{M}.$$

Because $\hat{\bar{Y}}$ is a ratio of two estimates, the well known approximation for the mean square error (Cochran 1977, pp. 32-33) can be used:

$$\begin{aligned} \text{MSE}(\hat{Y}) &= E\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{z}}\right)^2 \approx E\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{Z}}\right)^2 = \\ &= \frac{1}{\bar{Z}^2} [E(\bar{x}^2) - 2\bar{Y}E(\bar{z}\bar{x}) + \bar{Y}^2E(\bar{z}^2)] = \\ &= \frac{M^2}{n^2 \left(\sum_{j=1}^M z_j\right)^2} \left[E\left(\sum_{i=1}^n x_{K_i}\right)^2 - \right. \\ &\quad \left. 2\bar{Y}E\left(\sum_{i=1}^n x_{K_i} \sum_{i=1}^n z_{K_i}\right) + \bar{Y}^2E\left(\sum_{i=1}^n z_{K_i}\right)^2 \right]. \end{aligned} \quad (2.12)$$

The first expectation in (2.12) is simply (2.9). Next, using (2.1) on the middle term in (2.12) results in

$$\begin{aligned} E\left(\sum_{i=1}^n x_{K_i} \sum_{i=1}^n z_{K_i}\right) &= E\left(\sum_{i=1}^n x_{K_i} \frac{1}{s_{K_i}}\right) + \\ E\left(\sum_{i=1}^n \sum_{i'=1}^n x_{K_i} \frac{1}{s_{K_{i'}}}\right) &= \frac{n}{N} \sum_{k=1}^M \frac{y_k}{s_k} + E\left(\sum_{i=1}^n \sum_{i'=1}^n x_{K_i} \frac{1}{s_{K_{i'}}}\right). \end{aligned}$$

Using (2.7) and (2.9) yields,

$$\begin{aligned} E\left(\sum_{i=1}^n \sum_{i'=1}^n x_{K_i} \frac{1}{s_{K_{i'}}}\right) &= n(n-1)E\left(x_{K_i} \frac{1}{s_{K_{i'}}}\right) = \\ n(n-1) \sum_{[jk,j'k'] \in R^{**}} \sum \left[\left(\frac{y_k}{s_k} \frac{1}{s_{k'}} \right) \Pr\left(x_{K_i} = \frac{y_k}{s_k}, \frac{1}{s_{K_{i'}}} = \frac{1}{s_{k'}}\right) \right] &= \\ n(n-1) \sum_{[jk,j'k'] \in R^{**}} \sum \frac{y_k}{s_k} \frac{1}{s_{k'}} \left(\frac{1}{N(N-1)} s_{jk} s_{j'k'} \right) &= \\ \frac{n(n-1)}{N(N-1)} \sum_{[jk,j'k'] \in R^{**}} \sum \frac{y_k s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}}. \end{aligned}$$

Note that the double sum is over all admissible *ordered* pairs of arcs. Therefore,

$$\begin{aligned} E\left(\sum_{i=1}^n x_{K_i} \frac{1}{s_{K_i}}\right) + E\left(\sum_{i=1}^n \sum_{i'=1}^n x_{K_i} \frac{1}{s_{K_{i'}}}\right) &= \\ \frac{n}{N} \sum_{k=1}^M \frac{y_k}{s_k} + \frac{n(n-1)}{N(N-1)} \sum_{[jk,j'k'] \in R^{**}} \sum \frac{y_k s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}} &= \\ \frac{n}{N} \left(\sum_{k=1}^M \frac{y_k}{s_k} + \frac{(n-1)}{(N-1)} \sum_{[jk,j'k'] \in R^{**}} \sum \frac{y_k s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}} \right). \end{aligned}$$

Finally, similar to (2.1),

$$\begin{aligned} E\left(\sum_{i=1}^n z_{K_i}\right)^2 &= E\left(\sum_{i=1}^n \frac{1}{s_{K_i}}\right)^2 = \\ \frac{n}{N} \left(\sum_{k=1}^M \frac{1}{s_k} + \frac{2(n-1)}{(N-1)} \sum_{[jk,j'k'] \in R^{**}} \sum \frac{s_{jk}}{s_k} \frac{s_{j'k'}}{s_{k'}} \right). \end{aligned}$$

Substituting these expectations into equation (2.12) yields (2.11).

3. ESTIMATORS FOR MANY-TO-MANY FRAMES UNDER STRATIFIED RANDOM SAMPLING

3.1 Introduction

In this section we develop the estimators for a population count, mean, and total in the many-to-many frame case, when stratified random sampling is used. First, however, it is necessary to describe the sampling methodology under which these estimates are appropriate. Figure 3.1 provides an example that will be used throughout this section.

3.2 The Sampling Methodology

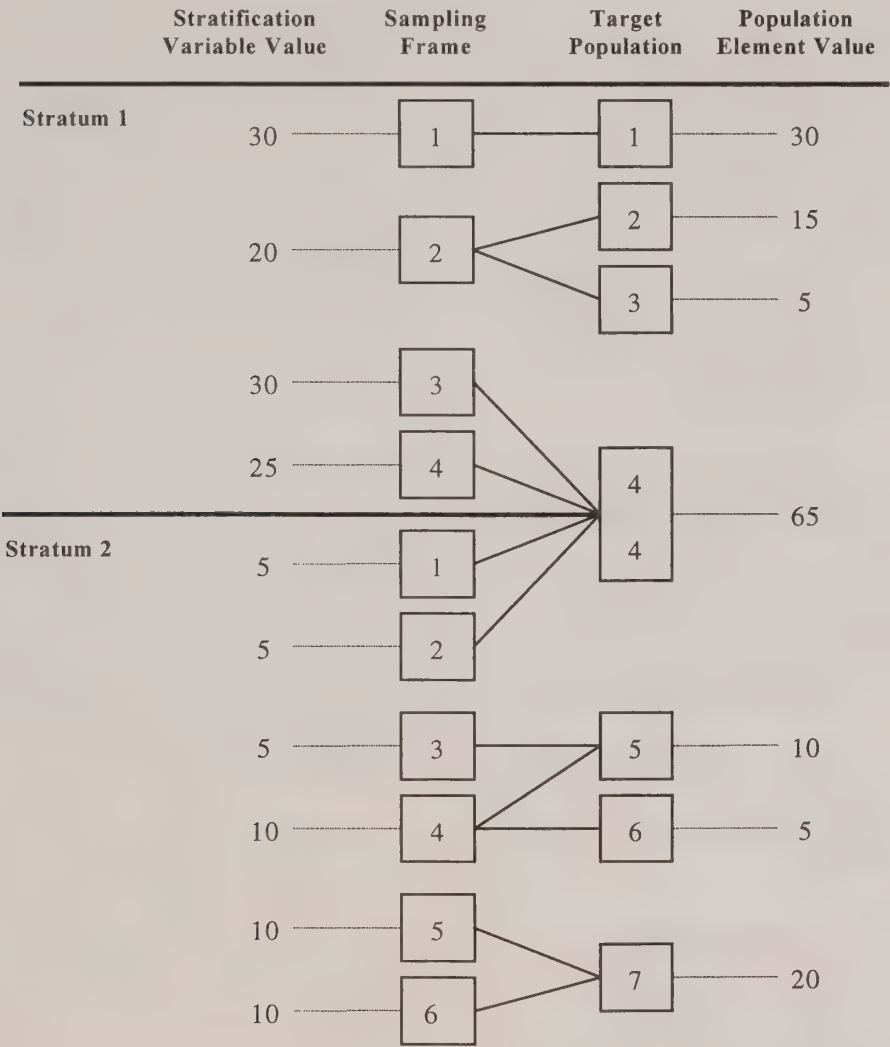
The same scenarios that were described in SRSWOR occur with respect to stratified random sampling. However, there are some additional problems that can arise in this case.

Consider the building characteristics study that motivated this research. Assume that the population element value in Figure 3.1 is the building size, and the stratification variable is electrical usage associated with the street address. Because the frame of street addresses had a many-to-many correspondence with the target population of commercial buildings, the following problems arose in addition to those mentioned in Section 2.1:

1. **Mis-stratification:** For example, frame unit (street address) 2 in stratum 1 appeared to be a large building because of the large electrical usage associated with it, and as a result, it was placed in the first stratum. The data collection revealed that the street address actually corresponded to two small buildings (population elements 2 and 3). In another example, frame units 5 and 6 in stratum 2 appeared to be two small buildings in the frame, and were placed in the second stratum. But, the corresponding population element 7 is one large building with two street addresses.
2. **Crossover:** For example, frame units 3 and 4 in stratum 1, and frame units 1 and 2 in stratum 2 each have a different street address and, as a result, appear in the frame to be two small and two large buildings. But, data

collection revealed that all four street addresses corresponded to only one building (*e.g.*, a strip mall). In this case, not only is mis-stratification a problem, but not all the frame units associated with a single building are included in the same strata. That is, one population element (*i.e.*, building) “crosses over” multiple strata.

In the next section we develop estimators for population totals and counts and show that these estimators are unbiased despite mis-stratification and crossover. As is usually the case, however, mis-stratification increases the variance of the estimates. Also, insofar as crossover induces mis-stratification, it too increases the variances of the estimates.



Note: Frame units were placed in stratum 1 if the value of the stratification variable was 20 or more. Otherwise, the frame units were placed in stratum 2.

Figure 3.1. An example of the correspondence between the frame and the target population in stratified random sampling

Table 3.1
Arc Probabilities for Figure 3.1

Arc hjk	1,1,1	1,2,2	1,2,3	1,3,4	1,4,4	2,1,4	2,2,4	2,3,5	2,4,5	2,4,6	2,5,7	2,6,7
s_{hjk}	1	1/2	1/2	1	1	1	1	1	1/2	1/2	1	1

3.3 Population Totals and Counts

3.3.1 Estimator for a Population Total

The estimator developed here involves a method of weighting which extends the estimator presented in Hansen *et al.* (1953a, pp. 62-64) to stratified random sampling when using a many-to-many frame.

Assume that F has been partitioned into L mutually exclusive and exhaustive strata F_1, \dots, F_L of size N_1, \dots, N_L respectively. Units in F_h will be denoted hj where $j = 1, \dots, N_h$ and $h = 1, \dots, L$. Also, assume that a stratified random sample (without replacement) of size $n = n_1 + \dots + n_L$ has been drawn, where n_h is the sample size from F_h . Let hJ_1, \dots, hJ_{n_h} denote random variables such that $hJ_i = hj$ if the i -th draw from F_h results in the selection of hj . Let hK_1, \dots, hK_{n_h} denote random variables such that $hK_i = k$ if the i -th draw from F_h is followed by the selection of k from T . If hjk denotes the arc that originates at frame unit hj in F_h and terminates at k in T , the marginal probability of the random arc (hJ_i, hK_i) is given by,

$$\Pr\{(hJ_i, hK_i) = (hjk)\} = \frac{1}{N_h} s_{hjk},$$

in which $s_{hjk} = \Pr(hK_i = k | hJ_i = hj)$ is an arc probability. Note that s_{hjk} is the conditional probability of selecting population element k in T given that frame unit hj has been chosen. Assuming equal randomization probabilities, Table 3.1 shows the arc probabilities for Figure 3.1.

Let W_k denote the set of frame units hj in F that have arcs with a destination at k in T . For example, $W_4 = \{(1, 3), (1, 4), (2, 1), (2, 2)\}$. Also, define the population element weight $s_k = \sum_{hj \in W_k} s_{hjk}$.

Table 3.2 contains the weights (s_k) for all the population elements in Figure 3.1. The same observations concerning arc weights (s_{hjk}) and population element weights (s_k) made in section 2.3.1 apply here.

Table 3.2
Population Element Weights (s_k) for Figure 3.1

k	1	2	3	4	5	6	7
s_k	1	1/2	1/2	1+1+1+1=4	1+1/2=3/2	1/2	1+1=2

For each $h = 1, \dots, L$ and $i = 1, \dots, n_h$, let x_{hK_i} be random variables such that $x_{hK_i} = y_k/s_k$ if k in T is selected as a result of the selection of some hj in F_h .

The estimator of a population total for stratified random sampling, when using a sampling frame with a many-to-many structure is:

$$\hat{Y}_{st} = \sum_{h=1}^L \hat{Y}_h, \text{ where } \hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} x_{hK_i}. \quad (3.1)$$

3.3.2 Variance of the Estimator for a Population Total

Prior to developing the variance of estimator (3.1), some additional terminology must be defined. Let q_{hk} denote the

“stratum element weight”. This additional weight is necessary because of the potential of crossover. Let U_{hk} denote the set of frame units in F_h that have arcs with a destination at population element k . For example, $U_{24} = \{(2, 1), (2, 2)\}$. Then define $q_{hk} = \sum_{hj \in U_{hk}} s_{hjk}$. To illustrate, recall in Figure 3.1 population element 4 is represented by two frame units in stratum 2, so $q_{24} = \sum_{2j \in U_{24}} s_{2j4} = 2$.

The weight q_{hk} plays the role of s_k when selection is restricted to F_h . In fact, $q_{hk} = s_k$ when there is no crossover. The probability of selecting any frame unit from F_h on step i out of n_h is $1/N_h$. But, the probability of selecting a population element k represented by a frame unit in F_h is $\Pr(hK_i = k) = q_{hk}/N_h$, for all $i = 1, \dots, n_h$.

In order to develop the proof in this section, we introduce the term “apportioned stratum total” denoted by Y_h^* .

In effect, the values of the population elements that are represented by frame units in multiple strata are apportioned among those strata. Let V_h denote the set of population elements associated with frame units in F_h . In our example $V_1 = \{1, 2, 3, 4\}$ and $V_2 = \{4, 5, 6, 7\}$. Let

$$Y_h^* = \sum_{k \in V_h} y_k q_{hk} / s_k$$

where y_k is the value of population element k , $k = 1, 2, \dots, M$. When crossover is present, use of the weights q_{hk} and s_k apportion the measure y_k among the strata in which population element k is represented. We can think of the use of these weights as distributing the population element value among the strata depending upon the number of times the population element is represented in a stratum relative to the total number of times it is represented in the frame. For example in Figure 3.1 Y_1^* and Y_2^* are calculated as follows:

$$Y_1^* = \frac{30(1)}{1} + \frac{15(1/2)}{1/2} + \frac{5(1/2)}{1/2} + \frac{65(2)}{4} = 82.5$$

$$Y_2^* = \frac{65(2)}{4} + \frac{10(3/2)}{3/2} + \frac{5(1/2)}{1/2} + \frac{20(2)}{2} = 67.5.$$

Note that $\sum_{h=1}^L Y_h^* = Y$ whether or not crossover exists.

Theorem 3-1: The estimator for a population total (3.1) is unbiased.

Proof:

From (3.1),

$$E(\hat{Y}_{st}) = \sum_{h=1}^L \frac{N_h}{n_h} \sum_{i=1}^{n_h} E(x_{hK_i}). \quad (3.2)$$

For each $i = 1, \dots, n_h$,

$$E(x_{hK_i}) = \sum_{k \in V_h} \frac{y_k}{s_k} \Pr(hK_i = k) =$$

$$\sum_{k \in V_h} \frac{y_k}{s_k} \frac{q_{hk}}{N_h} = \frac{1}{N_h} \sum_{k \in V_h} \frac{y_k q_{hk}}{s_k} = \frac{1}{N_h} Y_h^*. \quad (3.3)$$

Substituting (3.3) into equation (3.2) yields $E(\hat{Y}_{st}) = Y$.

In the main result below we need the following notation. Let R_h^* and R_h' be the set of admissible and inadmissible unordered pairs of arcs originating in F_h , respectively. Definitions of the above are identical to the corresponding concepts for the SRSWOR case, but restricted now to strata.

Theorem 3-2: The variance of (3.1) is:

$$V(\hat{Y}_{st}) = \sum_{h=1}^L S_h^2, \quad (3.4)$$

where,

$$S_h^2 = \frac{N_h}{n_h} \left[\sum_{k \in V_h} q_{hk} \left(\frac{y_k}{s_k} \right)^2 + \frac{2(n_h - 1)}{(N_h - 1)} \times \sum_{[hjk, hj'k'] \in R_h^*} \left(\frac{y_k s_{hjk}}{s_k} \frac{y_{k'} s_{hj'k'}}{s_{k'}} \right) \right] - \left(\sum_{k \in V_h} \frac{y_k q_{hk}}{s_k} \right)^2. \quad (3.5)$$

Proof: First write,

$$V(\hat{Y}_{st}) = E(\hat{Y}_{st})^2 - Y^2 = E\left(\sum_{h=1}^L \hat{Y}_h^2\right) - \sum_{h=1}^L (Y_h^*)^2 + 2\left(E\left(\sum_{h < h'} \hat{Y}_h \hat{Y}_{h'}\right) - \sum_{h < h'} Y_h^* Y_{h'}^*\right). \quad (3.6)$$

The last two terms cancel because \hat{Y}_h and $\hat{Y}_{h'}$ are independent. This follows since apportionment creates a new stratified population containing no crossover and samples chosen within different strata are independent. Thus, with

$$S_h^2 = E(\hat{Y}_h^2) - (Y_h^*)^2, \quad V(\hat{Y}_{st}) = E\left(\sum_{h=1}^L \hat{Y}_h^2\right) - \sum_{h=1}^L (Y_h^*)^2 = \sum_{h=1}^L S_h^2.$$

Now,

$$E(\hat{Y}_h^2) = \frac{N_h^2}{n_h^2} E\left(\sum_{i=1}^{n_h} x_{hK_i}\right)^2 = \frac{N_h^2}{n_h^2} \left(\sum_{i=1}^{n_h} E(x_{hK_i})^2 + 2E\left(\sum_{i < i'} x_{hK_i} x_{hK_{i'}}\right) \right). \quad (3.7)$$

For each $i = 1, \dots, n_h$,

$$E(x_{hK_i})^2 = \sum_{k \in V_h} \left[\left(\frac{y_k}{s_k} \right)^2 \Pr(hK_i = k) \right] = \sum_{k \in V_h} \left[\left(\frac{y_k}{s_k} \right)^2 \frac{q_{hk}}{N_h} \right]. \quad (3.8)$$

Then, using equation (2.7) and (2.8),

$$\begin{aligned} 2E\left(\sum_{i > i'} x_{hK_i} x_{hK_{i'}}\right) &= 2 \binom{n_h}{2} E(x_{hK_i} x_{hK_{i'}}) = \\ n_h(n_h - 1) \sum_{[hjk, hj'k'] \in R_h^*} \frac{y_k}{s_k} \frac{y_{k'}}{s_{k'}} \Pr(hK_i = k, hK_{i'} = k') &= \\ n_h(n_h - 1) \sum_{[hjk, hj'k'] \in R_h^*} \left[\left(\frac{y_k}{s_k} \frac{y_{k'}}{s_{k'}} \right) \left(\frac{N_h}{2} \right)^{-1} s_{jk} s_{j'k'} \right] &= \\ \frac{2n_h(n_h - 1)}{N_h(N_h - 1)} \sum_{[hjk, hj'k'] \in R_h^*} \left[\frac{y_k s_{hjk}}{s_k} \frac{y_{k'} s_{hj'k'}}{s_{k'}} \right]. \end{aligned} \quad (3.9)$$

Equation (3.5) now follows from (3.8), (3.9), and the definition of Y_h^* .

Using the method of Corollary 2-1, (3.5) can be simplified for computing purposes as follows:

$$\begin{aligned} S_h^2 &= \frac{N_h}{n_h} \left[\sum_{k \in V_h} q_{hk} \left(\frac{y_k}{s_k} \right)^2 + \frac{(n_h - 1)}{(N_h - 1)} \left[\left(\sum_{hjk \in A_h} \frac{y_k s_{hjk}}{s_k} \right)^2 - \right. \right. \\ &\quad \left. \left. \sum_{hjk \in A_h} \left(\frac{y_k s_{hjk}}{s_k} \right)^2 - 2 \sum_{[hjk, hj'k'] \in R_h'} \left(\frac{y_k s_{hjk}}{s_k} \frac{y_{k'} s_{hj'k'}}{s_{k'}} \right) \right] \right] - \\ &\quad \left(\sum_{k \in V_h} \frac{y_k q_{hk}}{s_k} \right)^2, \end{aligned}$$

where A_h denotes the set of arcs that originate at frame units in F_h .

3.4 Population Means

3.4.1 Estimator for a Population Mean

The estimator developed here for a population mean for stratified random sampling extends the estimator presented by Hansen *et al.* 1953a (pp. 62-64) to the case of a stratified random sample from a many-to-many frame.

The estimator for a population mean when using stratified random sampling and a many-to-many frame is:

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^L \frac{N_h}{N} \hat{\bar{Y}}_h, \quad \text{where } \hat{\bar{Y}}_h = \frac{\sum_{i=1}^{n_h} x_{hK_i}}{\sum_{i=1}^{n_h} \frac{1}{s_{hK_i}}}. \quad (3.10)$$

As in the SRSWOR case, the estimator for a population mean is biased because it is a ratio estimator.

4. CONCLUSIONS

In this paper we have developed estimators for population totals, counts and means that are appropriate when the sampling frame has a many-to-many structure. We have focused on simple random sampling and stratified random sampling designs.

We used the method of weighting described in this paper in a study of commercial buildings for which a stratified random sample was employed. In this study, for which the sampling frame consisted of street addresses, interviewers recorded any additional street addresses that pertained to the selected building. It was then determined whether or not these additional street addresses were listed in the sampling frame, and whether or not they were connected to other population elements (commercial buildings). In more complex scenarios, the interviewers sometimes resorted to schematic sketches of the buildings and labelling all the pertinent addresses. This allowed us to determine the structure of all MC subgraphs in our sample and to develop the appropriate weights s_k .

In addition, we developed formulas for the variance of some of the estimators presented in this paper. It should be noted that these variance formulas are population parameters and do not translate readily into corresponding sample estimates. In fact, the authors are unaware of any optimal method for estimating the variances discussed in this paper. However, there are many computer intensive methods (balanced repeated replication, bootstrapping, etc.) for estimating variances in complex sample surveys (Wolter 1985). It should be emphasized that when using our estimators, each of these variance estimation schemes aims at a common target: the variance formulas we have developed.

Nevertheless, the usefulness of these variance formulas is in their application to the task of exploring the effects of frame imperfections, along with population characteristics, on the precision of estimation. Such an exploration, another future area of research, should result in recommendations and guidelines for the survey researcher on how to manage a frame with a many-to-many structure. That is, based upon frame and population characteristics, the survey researcher would be able to make strategic decisions concerning the options available: canvassing a population to remove correspondence imperfections, or using the estimators described herein.

Another area of future research is a comparison of the precision of our estimators to that of other estimators, such as the Horvitz-Thompson estimator. As noted in the introduction the Horvitz-Thompson estimator can be applied to sampling involving a many-to-many frame structure. An advantage of the Horvitz-Thompson estimator is that with properly identified first and second order inclusion probabilities, one can obtain both an estimate of a population characteristic and an unbiased estimate of its variance. In addition, the first order inclusion probabilities can be derived in a manner similar to Musser (1993) based only upon information from the MC subgraphs. However, these probabilities are very difficult to

compute in a complex many-to-many frame structure such as ours. It is, however, relatively easy to calculate the necessary weights for our estimators.

ACKNOWLEDGMENTS

The authors thank the individuals at Cinergy Corporation for the opportunity to conduct the building characteristics survey which motivated this research. We also thank the three referees for their excellent suggestions that have led to a significant improvement of this paper.

REFERENCES

- BANDYOPADHYAY, S., and ADHIKARI, A.K. (1993). Sampling from imperfect frames with unknown amount of duplication. *Survey Methodology*, 19, 193-197.
- BIRNBAUM, Z.W., and SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, PHS Publication 1000, Ser. 2, *Data Evaluation and Methods Research*, no. 11. Hyattsville, MD: National Center for Health Statistics, Public Health Service, U.S. Department of Health and Human Services.
- CASADY, R.J., and SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-605.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: Wiley & Sons.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953a). *Sample Survey Methods and Theory 1, Methods and Applications*. New York: Wiley & Sons.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953b). *Sample Survey Methods and Theory 2, Theory*. New York: Wiley & Sons.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley & Sons.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- LESSLER, J.T., and KALSBECK, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley & Sons.
- MUSSER, O. (1993). Unbiased estimation in the presence of frame duplication. *Proceedings of the International Conference on Establishment Surveys*, 889-892.
- SIRKEN, M.G. (1972a). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- SIRKEN, M.G. (1972b). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 65, 224-227.
- U.S. DEPARTMENT OF ENERGY, Energy Information Administration (1992). *Commercial Buildings Energy Consumption Survey*.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WRIGHT, T., and TSAO, H.J. (1983). A frame on frames: An annotated bibliography, (Ed., Tommy Wright). *Statistical Methods and the Improvement of Data Quality*, Orlando, Florida: Academic Press, 25-72.

Optimal Recursive Estimation for Repeated Surveys

IBRAHIM S. YANSANEH and WAYNE A. FULLER¹

ABSTRACT

Least squares estimation for repeated surveys is addressed. Several estimators of current level, change in level and average level for multiple time periods are developed. The Recursive Regression Estimator, a recursive computational form of the best linear unbiased estimator based on all periods of the survey, is presented. It is shown that the recursive regression procedure converges; and that the dimension of the estimation problem is bounded as the number of periods increases indefinitely. The recursive procedure offers a solution to the problem of computational complexity associated with minimum variance unbiased estimation in repeated surveys. Data from the U.S. Current Population Survey are used to compare alternative estimators under two types of rotation designs: the intermittent rotation design used in the U.S. Current Population Survey, and two continuous rotation designs.

KEY WORDS: Recursive regression estimation; Composite estimation; Rotation designs; Rotation groups.

1. INTRODUCTION

We consider least squares estimation for surveys conducted on repeated occasions with partial overlap of sampling units. See Duncan and Kalton (1987) for a general discussion of different types of surveys and the objectives of such surveys. In this paper, we shall be concerned with rotating panel surveys, in which repeated determinations are made on some sampling units but not every unit appears in the sample at every time point.

Theoretical foundations for the design and estimation for repeated surveys based on generalized least squares procedures were laid down by Patterson (1950), following initial work by Cochran (1942) and Jessen (1942). Least squares procedures have been considered further by several other authors. See, for instance, Fuller (1990), and the references cited therein. Least squares estimation for a fairly general class of repeated surveys was considered by Yansaneh (1992). Composite estimation is a procedure of estimation for repeated surveys which makes use of the observations from the current and preceding periods, and the estimator of level from the preceding period. Breau and Ernst (1983) compared various alternative estimators to a composite estimator for the U.S. Current Population Survey (CPS). Kumar and Lee (1983) did similar work using data from the Canadian Labor Force Survey (LFS). Wolter (1979) provided a general composite estimation strategy for two-level rotation schemes such as the one used in the U.S. Census Bureau's Retail Trade Survey. Singh (1996) has proposed an alternative class of composite estimators. These authors assumed the unknown quantities on each occasion to be fixed parameters. Other authors, such as Scott, Smith, and Jones (1977), Jones (1980), Binder and Dick (1989), Bell and Hillmer (1990), and Pfeiffermann (1991) considered estimation for repeated surveys under the

assumption that the underlying true values constitute a realization of a time series.

In this paper, we discuss estimation procedures for repeated surveys, under the assumption that the unknown true values are fixed parameters. The estimators are compared to the method of composite estimation currently used in the CPS. The paper is organized as follows: In section 2, we state some basic assumptions regarding the general class of repeated surveys considered in this paper. A description of the CPS method of composite estimation is given in section 3. The method of best linear unbiased estimation is discussed in section 4. In section 5, we present a recursive regression estimation procedure designed to reduce the computational complexity associated with best linear unbiased estimation. Section 6 is devoted to an application to data from the CPS. Alternative estimators and rotation designs are compared.

2. BASIC ASSUMPTIONS

In this section, we describe surveys of the type we will study. A rotation group is a set of individuals selected for the sample and observed for a fixed number of periods and in a fixed pattern over time. Assume that in each period of the survey, s rotation groups are included in the sample, where $s > 1$ is fixed. Assume that the basic data from the survey can be organized in a set of elementary estimators (such as simple sample means and estimated totals) of the parameters of interest (such as population means and totals), where a set of elementary estimators is associated with each rotation group. For computational convenience, the data for p periods can be arranged in a $p \times s$ data matrix, denoted by H , in such a way that the observations on a rotation group appear in only one column. The total

¹ Ibrahim S. Yansaneh, Statistical Group, Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850; and Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50011 U.S.A.

number of elementary estimators is $n = p \times s$. We call the columns of H streams. Several rotation groups can appear in a stream. Assume that:

- (1) A given rotation group in a stream is observed over a period of total length $m + 1$, and the observation pattern for rotation groups is fixed and is the same for all groups.
- (2) The design is balanced on time-in-sample. That is, of the s rotation groups included in the sample at a given time, one group is being observed for the first time, one is being observed for the second time, ..., one is being observed for the last time, where the last time is separated by m periods from the first observation.

These assumptions are satisfied by surveys such as the CPS and the Canadian Labor Force Survey. See Yansaneh (1997) for an illustration of the 4-8-4 rotation scheme used in the CPS.

3. THE CPS COMPOSITE ESTIMATOR

In general, composite estimators combine recent estimator(s) and data from the current and preceding period(s) to form an estimator for the current period. With the CPS, six of the eight rotation groups observed at time t were observed at time $t - 1$. We shall refer to these six rotation groups as continuing rotation groups, and the remaining two as incoming rotation groups.

The composite estimator currently in use is determined by two parameters. The estimator is

$$\hat{\theta}_{t,c} = (1 - \pi_1)\bar{y}_t + \pi_1(\hat{\theta}_{t-1,c} + \hat{\delta}_{t,t-1}) + \pi_2\hat{\delta}_t \quad (1)$$

where, for the estimator currently used, $\pi_1 = 0.4$ and $\pi_2 = 0.2$, $y_{t,k}$ is the elementary estimate of level obtained from the rotation group which is in its k -th time in sample at time t , $\bar{y}_t = 8^{-1} \sum_{k=1}^8 y_{t,k}$ is the basic estimator, defined as the mean of the elementary estimates based on the eight rotation groups observed at time t , $\hat{\theta}_{t-1,c}$ is the composite estimator for time $t - 1$, $\hat{\delta}_{t,t-1}$ is an estimate of change in level, based on the six continuing rotation groups at time t , and $\hat{\delta}_t$ is the difference between the averages of the two incoming rotation groups and the six continuing rotation groups. Thus,

$$\hat{\delta}_{t,t-1} = 6^{-1} \sum_{k \in S} (y_{t,k} - y_{t-1,k-1}),$$

and

$$\hat{\delta}_t = 8^{-1} \left(\sum_{k \in T} y_{t,k} - 3^{-1} \sum_{k \in S} y_{t,k} \right),$$

where $T = \{1, 5\}$ and $S = \{2, 3, 4, 6, 7, 8\}$. The composite estimator used until 1985 contained only the first two terms on the right of (1). The third term was introduced for the

purpose of reducing the time-in-sample effects appearing in the original estimator. The incoming rotation groups produce larger estimates of unemployed than do the continuing rotation groups. Therefore, the direct difference $\hat{\delta}_{t,t-1}$ is influenced by the fact that the rotation group in its first time-in-sample has a larger expected value than that of the second time-in-sample. The time-in-sample effects do not cancel out in the difference estimate. The third term is an adjustment term which has the effect of reducing both the variance of the original composite estimator and the bias associated with time-in-sample effects. See Bailer (1975) or Breau and Ernst (1983) for a discussion of the bias of the pre-1985 composite estimator due to time-in-sample effects. We shall refer to the three-term composite estimator currently used in the CPS as the CPS Composite Estimator. This estimator has a variance close to that of the best linear unbiased estimator for monthly estimates of unemployment level. Let $y_{i,t}$, $i = 1, 2, \dots, s$, be the elementary estimator of the parameter of interest obtained from the rotation group which is in stream i at time t . The CPS composite estimator can be written as

$$\hat{\theta}_{t,c} = \sum_{i=1}^8 \omega_{1,k(i,t)} y_{i,t} + \sum_{i=1}^8 \omega_{2,k(i,t)} y_{i,t-1} + \pi_1 \hat{\theta}_{t-1,c} \quad (2)$$

where $k(i,t) = k$ defines the time-in-sample of observation (it) as a function of the stream (i) and time (t). If $\lambda_1 = 1/8$ and $\lambda_2 = -1/6$, and $\lambda_3 = 1/3$, then $\omega_{2,k} = \pi_1 \lambda_2$, and

$$\omega_{1,k} = \begin{cases} (1 - \pi_1)\lambda_2 - \pi_1\lambda_2 - \pi_2\lambda_1\lambda_3 & \text{for } k \in S \\ \lambda_1(1 - \pi_1 + \pi_2) & \text{for } k \in T \end{cases}$$

Let

$$P_1 = (\omega_{1,k(1,t)}, \omega_{1,k(2,t)}, \dots, \omega_{1,k(8,t)})',$$

$$P_2 = (\omega_{2,k(1,t)}, \omega_{2,k(2,t)}, \dots, \omega_{2,k(8,t)})',$$

and $y_t = (y_{1,t}, y_{2,t}, \dots, y_{8,t})'$. Then,

$$\hat{\theta}_{t,c} = P_1' y_t + P_2' y_{t-1} + \pi_1 \hat{\theta}_{t-1,c} \quad (3)$$

Substituting in (3) recursively, we have, for an estimator initiated at time zero,

$$\hat{\theta}_{t,c} = P_1' y_t + \sum_{j=1}^t \pi_1^{j-1} (P_2 + \pi_1 P_1)' y_{t-1} \quad (4)$$

Equation (4) is an expression of $\hat{\theta}_{t,c}$ as a linear function of current and past observations, where the weight of an observation declines as its distance from the current period increases.

4. BEST LINEAR UNBIASED ESTIMATION

Suppose $\Theta_p = (\theta_1, \theta_2, \dots, \theta_p)'$ is the $p \times 1$ vector of parameters of interest, where $\theta_t, t = 1, 2, \dots, p$, is the level of the parameter of interest at time t . Thus at time j , θ_j is the current level of the parameter of interest. For example, in the context of the CPS, θ_j might represent the population mean or proportion of unemployed at time j . Our objective is to construct efficient estimators of the current level of the parameters. The change in level and average level over multiple periods of time are also of interest.

The best linear unbiased estimator (BLUE) of the current level is defined to be the minimum-variance unbiased linear combination of the elementary estimators from the rotation groups available for estimation. It is possible in the process of computing the BLUE for the current level, to also compute the BLUEs for all periods using data available at the current time.

Suppose that a repeated survey has been in operation for p periods and that s streams of data collected over p periods are available for estimation. Let $y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,p})'$ be the vector of p observations in the i -th stream at time t . Let Y_p be the data vector formed by the streams or columns of the $p \times s$ data matrix H , arranged chronologically. Thus, $Y_p = (y_1', y_2', \dots, y_s')'$ is an $n \times 1$ vector of observations, where $n = s \times p$. Let $X = J_{s \times 1} \otimes I_{p \times p}$ be the $n \times p$ design matrix which relates the estimates in Y_p to their expected values in Θ_p ; where $J_{s \times 1}$ is the $s \times 1$ vector of ones, $I_{p \times p}$ is the identity matrix of order p , and \otimes denotes the Kronecker product. The linear model for Y_p is

$$Y_p = X\Theta_p + U_p \quad (5)$$

where U_p is the vector of error terms satisfying the assumptions $E(U_p) = 0$ and $E(U_p U_p') = V_p$, where V_p is assumed to be a known, symmetric, and positive definite matrix. By the Gauss-Markov Theorem, the BLUE of Θ_p is

$$\hat{\Theta}_p = (X' V_p^{-1} X)^{-1} X' V_p^{-1} Y_p.$$

The covariance matrix of $\hat{\Theta}_p$ is $\sum_p = (X' V_p^{-1} X)^{-1}$.

5. RECURSIVE REGRESSION ESTIMATION

Recursive estimation techniques have been found useful in situations where data do not all become available at the same time but rather accumulate over time, and the computation of optimal estimates based on all available data is impractical. See, for example, Odell and Lewis (1971), Sallas and Harville (1981) and references cited therein, for recursive algorithms for best linear unbiased estimation. Tiller (1989) presented a Kalman-filter approach to estimation of labor force characteristics using survey data.

As described in Section 4, the direct computation of the BLUE becomes progressively more complicated as the

number of periods increases. We develop a recursive regression estimation procedure for repeated surveys that uses a judiciously chosen set of initial estimates, new observations of the current level, and the previous observations on the currently observed rotation groups to produce the BLUE of current level.

5.1 Transformed Elementary Estimates and a Proposed Estimator

Suppose a survey has been in operation for at least m periods and assume:

- (3) The rotation groups are independent.
- (4) The covariance structure of the observations is known.
- (5) The covariance structure of the observations in a stream is constant over time, and it is the same for all streams.

These assumptions are used in the construction of a linear estimator. Assumption (3) will be relaxed for the computation of the variance of the estimator. Under assumptions (1) and (3), observations that are more than m periods apart are independent. At the current time, denoted by c , where $c > m$, a set of s elementary estimates of the parameter θ_c are observed. To construct the generalized least squares estimator, the s current observations are transformed so that they are uncorrelated with previous observations. After transformation, the expected values of the transformed observations are functions of θ_c and the parameters for the m preceding periods. Assume that the BLUE of the vector of parameters for the previous m periods, and the $m \times m$ covariance matrix of these estimators, are available. Thus, at time c , we have: (i) m initial estimates $\hat{\Theta}_{c-1(m)} = (\hat{\theta}_{c-m}, \dots, \hat{\theta}_{c-1})'$; (ii) the covariance matrix $\sum_{c-1(m)}$ of $\hat{\Theta}_{c-1(m)}$; and (iii) s independent observations on the s streams at the current time. Let the transformed observations, denoted by $z_{ic}, i = 1, 2, \dots, s$, be

$$z_{ic} = y_{i,c} - \sum_{j=1}^m b_{k(i,c),j} y_{i,c-j} \quad (6)$$

where $b_{k(i,c),j}$ are the coefficients such that $z_{i,c}$ is uncorrelated with $y_{i,c-j}$ for all $j > 0$. By assumptions (4) and (5), the coefficients $b_{k(i,c),j}$ are fixed over time. By assumption (3), $z_{i,c}$ is uncorrelated with all earlier observations. The expected value of $z_{i,c}$ is $\theta_c - \sum_{j=1}^m b_{k(i,c),j} \theta_{c-j}$, $i = 1, 2, \dots, s$.

5.2 The Recursive Regression Estimator

Let $\hat{\theta}_h(t)$, $h \leq t$, denote the least squares estimator of the (scalar) parameter θ_h constructed using data through time t ; and let $\hat{\Theta}_{t(m)} = (\hat{\theta}_{t-m+1}(t), \dots, \hat{\theta}_t(t))'$ denote the least squares estimator of the vector of m parameters $\theta_{t-m+1}, \dots, \theta_t$, at time t constructed using data through time t . Our objective is to construct the minimum variance

estimator for θ_c , the current level of the parameter of interest using all data available at time c . A linear model for data available at the current time is

$$\mathbf{Z}_c = \mathbf{W}\Theta_{c(m+1)} + \mathbf{U}_c \quad (7)$$

where

$$\mathbf{W} = \begin{pmatrix} \mathbf{I}_m & \mathbf{0} \\ \mathbf{X}_{21} & \mathbf{J}_s \end{pmatrix},$$

$\mathbf{Z}_c' = (\hat{\Theta}_{c-1(m)}', \mathbf{z}_c')$, $\mathbf{z}_c' = (z_{1c}, \dots, z_{sc})$, and \mathbf{X}_{21} is an $s \times m$ matrix whose entries are constant over time, and are functions of the coefficients $b_{k,j}$ of (6). If $\text{Var}\{z_{i,c}\} = \sigma_i^2$, $i = 1, 2, \dots, s$, and Ω_{22} is the diagonal matrix with σ_i^2 as the diagonal entries, then the covariance matrix of \mathbf{Z}_c is $\mathbf{V}_c = \text{blockdiag}\{\sum_{c-1(m)}, \Omega_{22}\}$. It is assumed that σ_i^2 , $i = 1, 2, \dots, s$, are positive.

The recursive regression estimator (RRE) of $\Theta_{c(m+1)}$ is defined to be the least squares estimator of $\Theta_{c(m+1)}$ based on model (7). Thus the RRE of $\Theta_{c(m+1)}$ is

$$\hat{\Theta}_{c(m+1)} = (\mathbf{W}' \mathbf{V}_c^{-1} \mathbf{W})^{-1} \mathbf{W}' \mathbf{V}_c^{-1} \mathbf{Z}_c \quad (8)$$

and the covariance matrix of $\hat{\Theta}_{c(m+1)}$ is $\mathbf{Q}_{c(m+1)} = (\mathbf{W}' \mathbf{V}_c^{-1} \mathbf{W})^{-1}$.

The utility of the estimator (8) is its computational simplicity. At any fixed time t in a repeated survey, all the information relevant to the problem of estimating $\theta_t, \theta_{t-1}, \dots, \theta_{t-m}$ can be obtained from a set of m recursive least squares estimates and the current observations.

We now describe more fully the recursive regression procedure. At time t , we have $\hat{\Theta}_{t(m+1)}$, the RRE of $\Theta_{t(m+1)}$, and its $(m+1) \times (m+1)$ covariance matrix $\sum_{t(m+1)}$. Partition $\sum_{t(m+1)}$ as

$$\sum_{t(m+1)} = \begin{pmatrix} v_{11,t} & \mathbf{V}_{12,t} \\ \mathbf{V}_{12,t}' & \sum_{t(m)} \end{pmatrix},$$

where $v_{11,t}$ is the variance of $\hat{\theta}_{t-m}(t)$, $\sum_{t(m)}$ is the covariance matrix of $(\hat{\theta}_{t-m+1}(t), \dots, \hat{\theta}_t(t))'$, and $\mathbf{V}_{12,t}$ is the covariance between these two quantities. Observe that if θ_{t-m} is retained in the parameter vector and $\hat{\theta}_{t-m}$ is retained in the data vector, the estimator of θ_{t+1} is unchanged (the estimator of θ_{t-m} would, in general, be changed). This is because the estimator of the original parameter vector of a least squares problem is not changed if an observation whose expectation is equal to a single new parameter is added to the problem. Thus, to update the RRE for the next period, we drop the initial estimate for the earliest period, $\hat{\theta}_{t-m}(t)$, from the data vector, and drop the corresponding parameter θ_{t-m} from the parameter vector. The parameter θ_{t+1}

is then added to the parameter vector. In this way, the dimension of the basic model matrix \mathbf{W} of the estimation problem is kept constant over time. Thus in the class of repeated surveys considered in this paper, there is an upper bound on the computational effort required for the BLUE of the vector of parameters of interest.

The model at time $t+1$ may be written as model (7), with $c = t+1$, $\mathbf{Z}_{t+1} = (\hat{\theta}_{t-m+1}(t), \dots, \hat{\theta}_{t-1}(t), \hat{\theta}_t(t), \mathbf{z}_{t+1}')'$, $\Theta_{t+1(m+1)} = (\theta_{t-m+1}, \dots, \theta_t, \theta_{t+1})'$, and the covariance matrix of \mathbf{Z}_{t+1} is $\mathbf{V}_{t+1} = \text{blockdiag}\{\sum_{t(m)}, \Omega_{22}\}$. The BLUE of $\Theta_{t+1(m+1)}$ and its covariance matrix are then obtained from the usual least squares formulas. The least squares estimators of the last m elements of $\Theta_{t+1(m+1)}$ are then used as the initial estimates in the model for the next iteration.

The following theorem states that the covariance matrix of the vector of recursive least squares estimators converges to a positive definite matrix as the number of periods in the survey increases indefinitely. A proof is given in the appendix.

Theorem: At any time t , let the vector of recursive least squares estimators $\hat{\Theta}_{t(m)} = (\hat{\theta}_{t-m+1}(t), \dots, \hat{\theta}_{t-1}(t), \hat{\theta}_t(t))'$ be the BLUE of the vector of parameters $\Theta_{t(m)} = (\theta_{t-m+1}, \dots, \theta_{t-1}, \theta_t)'$ based on data through time t . Let $\sum_{t(m)}$ be the covariance matrix of $\hat{\Theta}_{t(m)}$. Let the assumptions (1) through (5) hold. Also assume that the elements of \mathbf{V}_n^{-1} are bounded for all n , where \mathbf{V}_n is the covariance matrix of any n observations. Then, the covariance matrix $\sum_{t(m)}$ converges as $t \rightarrow \infty$; and the limit is an $m \times m$ positive definite matrix.

6. APPLICATION TO THE U.S. CURRENT POPULATION SURVEY

6.1 The CPS Design

The CPS is a monthly household survey conducted by the United States Census Bureau in cooperation with the Bureau of Labor Statistics for the purpose of providing national estimates of labor force characteristics such as the number employed, unemployed, and in the civilian labor force; and other characteristics of the non-institutionalized civilian population. The sample design of the CPS contains a rotation scheme that includes the replacement of a fraction of the households in the sample each month. For any given month, the sample consists of eight time-in-sample panels or rotation groups, of which one is being interviewed for the first time, one is being interviewed for the second time, ..., and one is being interviewed for the eighth time. In other words, the interview scheme is balanced on time-in-sample. Households in a rotation group are interviewed for four consecutive months, dropped for the next eight succeeding months, and then interviewed for another four consecutive months. They are then dropped from the sample entirely. This system of interviewing is called the 4-8-4 rotation scheme, and is a special case of schemes described by Rao and Graham (1964).

6.2 Estimation and Variance Estimation Procedures

We use estimates of the covariance structure of data from the CPS to compare alternative estimators and rotation designs. See Adam and Fuller (1992) and Fuller, Adam and Yansaneh (1993) for a detailed description of the construction of the model, the estimation of its parameters, and the estimation of the covariance structure of observations within a given rotation group for various characteristics of interest. Because the rotation groups come from the same set of primary sampling units, they are not independent and a component is included in the covariances to reflect the fact that the primary sampling units do not change. The RRE is computed with the eight current simple estimators and the 15 estimators for the 15 preceding periods. In computing the RRE, the covariances are used to create eight linear combinations of the current and the preceding fifteen observations that are uncorrelated with the preceding fifteen observations. Because of the primary sampling unit effect, these linear combinations are correlated with observations more than 15 periods in the past and in the same stream. Hence, they are correlated with the preceding estimators. The correlations with earlier estimators, $\hat{\theta}_{t-i}$, $i = 1, \dots, 15$, are included in the covariance matrix when the estimator of θ_t is constructed. However, because only the most recent 15 observations are used, the resultant estimator of θ_t is not the BLUE for current level. The calculated covariance matrix of $(\hat{\theta}_{t-15}, \dots, \hat{\theta}_{t-1}, \hat{\theta}_t)'$ is correct and, because the primary sampling unit effect is modest, it is felt that the estimator has efficiency close to that of the BLUE.

We shall restrict attention to the estimation of various parameters for two characteristics of interest: Employed and Unemployed. For each characteristic, the parameters of interest are the current level and period-to-period change for up to 12 periods. The estimators considered for comparison are the CPS composite estimator; the RRE; and the BLUEs using 2, 3, 12, 16, and 24 periods, where the BLUE for p periods at time t is the least squares estimator constructed using data from time $t - p + 1$ through time t . Results are reported for BLUEs based on 12 and 16 periods. In following the practice of the U.S. Bureau of Labor Statistics for CPS estimators, the estimators are not modified as new data become available. Thus the estimator of change in level of a characteristic of interest between times $t - 1$ and t is not the best possible estimator given all available data. It is the difference between the best estimator at time t based on data through time t and the best estimator at time $t - 1$ based on data through time $t - 1$.

We do not consider seasonal adjustment in this discussion. However, the estimation procedures presented can be extended to include seasonal adjustments. To compute the variance of a given estimator at a given time, the estimator is first expressed as a linear combination of all the observations available at that time. The variance of

the estimator is then computed as a function of the coefficients of the linear combination and the entries of the covariance matrix.

6.3 Numerical Results and Discussion

6.3.1 Comparison of Alternative Estimators

The variances of the alternative estimators relative to the variance of the basic estimator of current level, for each of the characteristics of interest, are presented in Table 1. Recall that the basic estimator of the current level, denoted by \bar{y}_t , is the simple mean of the eight elementary estimators obtained from the eight rotation groups observed at time t . That is, $\bar{y}_t = 8^{-1} \sum_{k=1}^8 y_{t,k}$, and $\text{Var}(\bar{y}_t) = \sigma^2/8$, where $\sigma^2 = \text{Var}(y_{t,k})$ for all t and k . The basic estimator of change between two periods is the difference between the simple means for the two periods.

The BLUE procedure based on 3 periods or more produces more efficient estimators of current level than the CPS composite estimator. In general, the best linear unbiased estimation procedure becomes more statistically efficient as the number of periods increases. For both characteristics, the results reveal that the best linear unbiased procedure based on 12 periods is uniformly more efficient than the CPS composite estimator for all parameters, except one-period change in unemployed. Recall that the estimator of change is not BLUE because the estimator is the difference of estimators constructed at time t and at time $t - 1$. Thus, the estimator called "BLUE" is best only for current level using the stated amount of data. The difference between the variance of the composite estimator of one-period change and the variance of the 12-period BLUE of one-period change in unemployed is less than one percent. The gain in precision of the best linear unbiased estimation procedure for employed relative to the CPS composite estimator for current level is 22% for the BLUE for 12 periods, 28% for the BLUE for 16 periods, 30% for the BLUE for 24 periods, and 33% for the RRE. The corresponding gains for unemployed are 2%, 3%, and 3%. These results are a reflection of the nature of the autocorrelation functions of the characteristics. The autocorrelation function for unemployed declines much faster than that for employed.

With the exception of one-period change in employed, there is an improvement in the efficiency of the estimation of change from using the alternative estimators instead of the CPS composite estimator. The gain in precision increases as the number of periods increases, reaching a maximum value at five-period change for both characteristics. The gain then decreases slightly. In the case of the RRE, the maximum gain in efficiency for estimated change is 64% for employed and 5% for unemployed.

Table 1
Variances of alternative estimators relative to the variance of the basic estimator of current level

Parameter	Employed				Unemployed			
	CPS Comp.	BLUE for 12 periods	BLUE for 16 periods	Recursive Regression Estimator	CPS Comp.	BLUE for 12 Periods	BLUE for 16 periods	Recursive Regression Estimator
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Current Level	0.862	0.704	0.672	0.650	0.947	0.924	0.918	0.918
1-period change	0.511	0.457	0.437	0.432	1.070	1.077	1.073	1.073
2-period change	0.813	0.646	0.613	0.604	1.361	1.345	1.338	1.338
3-period change	1.065	0.763	0.724	0.711	1.528	1.481	1.473	1.473
4-period change	1.279	0.830	0.800	0.784	1.645	1.569	1.563	1.562
5-period change	1.363	0.880	0.847	0.829	1.691	1.614	1.607	1.606
6-period change	1.390	0.910	0.873	0.855	1.708	1.637	1.628	1.628
7-period change	1.388	0.930	0.884	0.865	1.710	1.646	1.637	1.636
8-period change	1.353	0.932	0.884	0.860	1.701	1.645	1.635	1.634
9-period change	1.255	0.912	0.854	0.832	1.671	1.624	1.614	1.614
10-period change	1.154	0.895	0.824	0.806	1.641	1.606	1.595	1.595
11-period change	1.061	0.883	0.795	0.782	1.614	1.590	1.578	1.578
12-period change	0.992	0.883	0.767	0.761	1.593	1.577	1.563	1.563

6.3.2 Comparison of Alternative Estimators and Rotation Designs

The variances of alternative estimators under various rotation designs are given in Table 2. All variances are relative to the variance of the basic estimator of current level under that design. The efficiencies of alternative estimators of current level, change in level, and average level for multiple time periods are compared under the intermittent 4-8-4 rotation design and two continuous rotation designs. The continuous rotation designs are the 6-continuous scheme and the 8-continuous scheme. The 6-continuous scheme is the rotation scheme used in the Canadian Labor Force Survey conducted by Statistics Canada. For each period of the survey, the sample consists of six rotation groups, one rotation group in its first time-in-sample, ..., and one rotation group in its sixth time-in-sample. A given rotation group remains in the sample for six consecutive periods and then permanently drops out of the sample. See Kumar and Lee (1983) for more details about the design of the Canadian Labor Force Survey. In the 8-continuous scheme, there are 8 rotation groups in the sample for each period. A given rotation group remains in the sample for eight consecutive periods and then permanently drops out of the sample.

We compare the performance under the various rotation designs using the BLUE of current level based on 36 periods. We call this estimator the “best estimator” because its efficiency is virtually the same as that of the RRE. For all rotation schemes under consideration, there are some improvements in the precision of the estimators of current level from using the best estimator relative to the CPS composite estimator. As seen in Table 2, the gain is highest for employed where, under the 4-8-4 rotation scheme, the variance of the best estimator of current level is only 76% of that of the CPS composite estimator.

The precision of the estimators of change relative to the precision of the CPS composite estimator depends on the rotation design. From Table 2, we see that under the 4-8-4 rotation scheme, there is some gain in precision, which increases as the lag increases. For employed, the variance of the least squares estimator is 85% of the variance of the CPS composite estimator for one-period change, 61% of the variance of the CPS composite estimator for six-period change, and 76% of the variance of the CPS composite estimator for 12-period change. (Compare columns (2) and (3) of Table 2.)

Table 2

Variances of alternative estimators and rotation designs; the variance of the basic estimator of current level under each design equals one

Parameter	Employed				Unemployed			
	CPS Comp.	Best Est. (4-8-4)	Best Est. (8 Cont)	Best Est. (6 Cont)	CPS Comp.	Best Est. (4-8-4)	Best Est. (8 Cont)	Best Est. (6 Cont)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Current Level	0.862	0.653	0.761	0.759	0.947	0.918	0.944	0.938
1-period change	0.511	0.432	0.395	0.434	1.070	1.073	1.003	1.051
2-period change	0.813	0.604	0.559	0.619	1.361	1.338	1.250	1.312
3-period change	1.065	0.710	0.669	0.747	1.528	1.473	1.372	1.443
4-period change	1.279	0.783	0.731	0.829	1.645	1.562	1.473	1.543
5-period change	1.363	0.828	0.782	0.901	1.691	1.606	1.533	1.607
6-period change	1.390	0.854	0.828	0.970	1.708	1.628	1.577	1.655
7-period change	1.388	0.863	0.874	1.026	1.710	1.636	1.612	1.686
8-period change	1.353	0.858	0.828	0.960	1.701	1.934	1.642	1.705
9-period change	1.255	0.830	0.960	1.108	1.671	1.614	1.663	1.719
10-period change	1.154	0.803	0.993	1.139	1.641	1.595	1.678	1.727
11-period change	1.061	0.779	1.021	1.165	1.614	1.578	1.688	1.733
12-period change	0.992	0.758	1.046	1.186	1.593	1.564	1.696	1.737
12-period average	0.369	0.326	0.440	0.394	0.255	0.249	0.301	0.266
12-change in averages	0.248	0.162	0.365	0.403	0.273	0.262	0.372	0.359

For estimating 12-period averages in employed using the 4-8-4 design, the CPS composite estimator is about 13% less efficient than the least squares estimator and, for estimating change in 12-period averages, it is about 53% less efficient, as can be seen by comparing the second and third columns of Table 2. For unemployed and the 4-8-4 design, there are only modest gains in precision from using the least squares estimator relative to the CPS composite estimator, as shown in the sixth and seventh columns of Table 2.

For estimation of 12-period change, 12-period average and change in 12-period averages, the 4-8-4 design is much superior to both continuous rotation designs for both characteristics. The continuous designs are generally superior for period-to-period changes for short periods.

6.3.3 Internal Consistency

In our analysis, we have constructed the best estimator of employed using only the past history of employed and the best estimator of unemployed using only the past history of

unemployed. There is no formal reason not to include the past history of both employed and unemployed in the construction of the estimators. However, Fuller *et al.* (1993) state that the estimated cross correlations are less than 0.10, suggesting that there is little gain from such inclusion.

A method of constructing estimates of multiple characteristics that are internally consistent was suggested by Fuller (1990). In this procedure, estimates of employed, unemployed, and not-in-the-labor-force are constructed. Then these estimates are used as controls in a regression procedure to construct weights for the current observations. The weights can then be used to construct internally consistent estimates of any parameter of interest. The estimation procedure, including estimates of subdivisions of the labor force, is planned for implementation in 1998 for the CPS. See Lent, Miller and Cantwell (1996).

6.4 Conclusions

The main conclusions emerging from the variance computations in this section can be summarized as follows:

1. For all rotation designs and all characteristics under consideration, there are alternative estimation procedures with a variance of the current level smaller than that of the CPS composite estimator.
2. For estimation of change under the 4-8-4 rotation design, the gain in precision of the alternative estimators relative to the CPS composite estimator increases as the lag increases, and peaks around the lag of minimum overlap.
3. The intermittent 4-8-4 rotation design is inferior to the continuous rotation designs for short-period changes, but superior for current level, long-period averages, and changes in long-period averages.
4. The CPS composite estimator is comparable to the RRE for unemployed for the estimation of one-period change and 12-period change. However, the recursive regression estimation procedure is superior to the CPS composite estimator for other measures of change.
5. The RRE is more efficient in estimating change in level at lags for which the CPS composite estimator is not targeted, for instance, lags of four months to six months.

ACKNOWLEDGMENTS

The authors thank the referee and John Eltinge for helpful comments on earlier versions of this paper. Research for this paper was partly supported by Joint Statistical Agreement 91-21 with the U.S. Bureau of the Census and Cooperative Agreement 43-3AEU-80088 with the National Agricultural Statistics Service and the U.S. Bureau of the Census. The views expressed in this paper are those of the authors and do not necessarily represent the policies of the Bureau of the Census, the Bureau of Labor Statistics, or the National Agricultural Statistics Service.

APPENDIX

Lemma 1. Let the assumptions of the theorem hold. Then the variance of the estimator of current level θ_c converges to a positive number as the number of periods increases.

Proof. If the means $\theta_{c-1}, \theta_{c-2}, \dots, \theta_{c-m}$ were known, then $g_{1c}, i = 1, 2, \dots, s$ are unbiased estimators of θ_c , where $g_{1c} = y_{1c}; g_{2c} = y_{2c} - b_{21}(y_{2,c-1} - \theta_{c-1}); \dots$; and $g_{sc} = y_{sc} - \sum_{j=1}^m b_{sj}(y_{s,c-j} - \theta_{c-j})$. Furthermore, $g_{1c}, i = 1, 2, \dots, s$ are independent with variances $\sigma_i^2, i = 1, 2, \dots, s$. We may write the linear model:

$$\mathbf{g} = \mathbf{J}_s \theta_c + \mathbf{e} \quad (\text{A1})$$

where $\mathbf{g} = (g_{1c}, g_{2c}, \dots, g_{sc})'$, \mathbf{J}_s is the $s \times 1$ column vector of ones, and \mathbf{e} is the $s \times 1$ vector of errors with $E(\mathbf{e}) = 0$, and $E(\mathbf{e}\mathbf{e}') = \mathbf{V}_s = \text{Diag}\{\sigma_1^2, \sigma_2^2, \dots, \sigma_s^2\}$. Thus the BLUE of θ_c for model (A1) has variance $(\sum_{i=1}^s \sigma_i^{-2})^{-1}$. By assumption,

the variances $\sigma_i^2, i = 1, 2, \dots, s$ are bounded below and the quantity $(\sum_{i=1}^s \sigma_i^{-2})^{-1}$ is a positive lower bound for the variance of the estimator of θ_c [see Lemma 4.2.3 of Yansaneh (1992)]. The variance of the estimator of θ_c is non-increasing as the number of observations increases, and hence, the variance converges to a positive number.

Lemma 2. Let the assumptions of the theorem hold. Then the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1}, \dots, \theta_{t-1}$, based on data through time t , converges to a positive number as t increases.

Proof. First, suppose at a fixed time τ , at least m periods of observations are available both prior to τ and after τ . Define a transformation of the following form for the observations in each of the s streams at time τ : $u_{i\tau} = y_{i\tau} - \sum_{j=-m}^m b_{k(i,\tau),j} y_{i,\tau-j}$, where $b_{k(i,\tau),0} = 0$ and $u_{i\tau}$ is uncorrelated with all observations preceding and succeeding $y_{i\tau}$ in the i -th stream. Let the variance of $u_{i\tau}$ be $\lambda_i^2, i = 1, 2, \dots, s$. These variances are bounded below by assumption. We conclude, as before, that there is a positive lower bound for the diagonal elements of the covariance matrix of the vector of recursive least squares estimators.

Now, assume that at time t , we begin the sequence of estimation with the vector of recursive least squares estimators $\hat{\Theta}_{t-1(m)} = (\hat{\theta}_{t-m}, \dots, \hat{\theta}_{t-1})'$ based on data for the preceding m periods; and the vector of transformed observations $\mathbf{z}'_t = (z_{1t}, \dots, z_{st})$. Thus the linear model for the data at time t is given by (7), with c replaced by t . The data vector \mathbf{Z}_t is of fixed dimension. Therefore, the covariance matrix of the BLUE of the vector of parameters $\Theta_{t(m+1)} = (\theta_{t-m}, \dots, \theta_{t-1}, \theta_t)'$ is $\sum_{t(m+1)} = (\mathbf{W}' \mathbf{V}_t^{-1} \mathbf{W})^{-1}$. For computational convenience, we express \mathbf{W} as $(\mathbf{I}_{m+1}, \mathbf{M})'$, where \mathbf{I}_{m+1} is the identity matrix of order $m+1$, and \mathbf{M} is an $(s-1) \times (m+1)$ matrix which is constant over time. Thus we have

$$\begin{aligned} \sum_{t(m+1)} &= (\Omega_{t-1(m+1)}^{-1} + \mathbf{M}' \Omega_{00}^{-1} \mathbf{M})^{-1} \\ &= \Omega_{t-1(m+1)} - \Omega_{t-1(m+1)} \mathbf{M}' \mathbf{D}_t^{-1} \mathbf{M} \Omega_{t-1(m+1)} \end{aligned} \quad (\text{A2})$$

where

$\Omega_{t-1(m+1)} = \text{blockdiag}\{\sum_{t-1(m)}, \sigma_1^2\}$, $\Omega_{00} = \text{diag}\{\sigma_2^2, \dots, \sigma_s^2\}$, and $\mathbf{D}_t = \Omega_{00} + \mathbf{M} \Omega_{t-1(m+1)} \mathbf{M}'$. Since the second term on the right hand side of (A2) is positive definite, we conclude that the first m diagonal elements of $\sum_{t(m+1)}$ are less than or equal to the original diagonal elements of $\sum_{t-1(m)}$. This means that as t increases, the variances of the estimators of $\theta_{t-m}, \dots, \theta_{t-2}, \theta_{t-1}$ are non-increasing. Since these variances are bounded below by a positive quantity, we conclude that the variances of the estimators of $\theta_{t-m}, \dots, \theta_{t-2}, \theta_{t-1}$ converge to positive numbers as t increases.

Lemma 3. Let the assumptions of the theorem hold. Then, the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, \dots, \theta_t - \theta_{t-1}$, based on data through time t , converges to a positive number as t increases.

Proof. First, we show that variance of the least squares estimator of $\theta_c - \theta_{c-1}$ (where c denotes the current period) converges as the number of periods increases by mimicking the arguments in the proof of Lemma 1. Also, arguments similar to those in the proof of Lemma 2 can be used to show that the variances of the least squares estimators of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, \dots, \theta_t - \theta_{t-1}$, all converge as the number of periods increases.

Proof of theorem. Since $\sum_{t(m)}$ is a submatrix of the covariance matrix $\sum_{t(m+1)}$ of the least squares estimators of the full set of parameters $\theta_{t-m}, \theta_{t-m+1}, \dots, \theta_{t-1}, \theta_t$, at time t , it is enough to show that $\sum_{t(m+1)}$ converges to a positive definite matrix as $t \rightarrow \infty$. From Lemma 1 and Lemma 2, each of the diagonal elements of $\sum_{t(m+1)}$ converges to a positive number as $t \rightarrow \infty$. From Lemma 3, the variance of the least squares estimator of each of the parameters $\theta_{t-m}, \theta_{t-m+1} - \theta_{t-m}, \dots, \theta_t - \theta_{t-1}$, converges to a positive number as $t \rightarrow \infty$. It follows that for each $j, 1 \leq j \leq m$, the covariance between the least squares estimators of θ_t and θ_{t-j} converges as $t \rightarrow \infty$ and hence the covariance matrix $\sum_{t(m+1)}$ converges as $t \rightarrow \infty$.

Next, we prove that the limiting covariance matrix is positive definite. Let $\lim_{t \rightarrow \infty} \sum_{t(m)} = \sum_{(m)}$. It is enough to show that the variance of any non-trivial linear combination of the recursive least squares estimators $\hat{\theta}_{t-j}(t)$, $j = 1, 2, \dots, m$, is bounded below by a positive quantity. Let v_{mm} be the lower bound of every linear combination of the observations with one of the coefficients equal to one. The bound is positive by the assumption that the elements of V_n^{-1} are bounded.

Now, every estimator of the parameter θ_{t-j} , $j = 0, 1, \dots, m$ is a linear combination of all observations such that the sum of the coefficients for the observations in the s streams at time $t-j$ is one, and the sum of the coefficients for the observations in the s streams at any other time is zero. This is a condition for the unbiasedness of the estimator for time $t-j$. For the sum of the coefficients of the s observations at time $t-j$ to be equal to one, at least one of the coefficients must be greater than or equal to s^{-1} . The minimum variance of any linear combination with first coefficient equal to s^{-1} is $s^{-2}v_{mm}$. Therefore, for $j = 0, 1, \dots, m$, $\text{Var}\{\hat{\theta}_{t-j}(t)\} \geq s^{-2}v_{mm}$.

Now, consider an arbitrary, non-trivial linear combination of the recursive least squares estimators $\hat{\theta}_{t-j}(t)$, $j = 0, 1, \dots, m$, given by $\sum_{j=0}^m \gamma_j \hat{\theta}_{t-j}(t)$, where, without loss of generality, $\gamma_0 = 1$. This linear combination can be expressed as

$$\sum_{j=0}^m \gamma_j \hat{\theta}_{t-j}(t) = \hat{\theta}_t(t) + \sum_{j=1}^m \gamma_j \hat{\theta}_{t-j}(t) \quad (\text{A3})$$

$$= \sum_{i=1}^s \sum_{h=1}^t c_{ih} y_{i,h} + \sum_{j=1}^m \gamma_j \sum_{i=1}^s \sum_{h=1}^t f_{ih(t-j)} y_{i,h}$$

$$= \sum_{i=1}^s \left[c_{ii} + \sum_{j=1}^m \gamma_j f_{ii(t-j)} \right] y_{i,t} + \sum_{i=1}^s \sum_{h=1}^{t-1} \left[c_{ih} + \sum_{j=1}^m \gamma_j f_{ih(t-j)} \right] y_{i,h}$$

where $c_{ii}, i = 1, 2, \dots, s$, are the coefficients of $y_{i,t}$ in $\hat{\theta}_t(t)$, and $f_{ih(t-j)}, j = 1, \dots, m$, are the coefficients of $y_{i,t}$ in $\hat{\theta}_{t-j}(t)$, $j = 1, \dots, m$, respectively. Therefore, $\sum_{i=1}^s c_{ii} = 1$, and $\sum_{i=1}^s f_{ih(t-j)} = 0$, for $j = 1, \dots, m$. Thus $\sum_{i=1}^s [c_{ii} + \sum_{j=1}^m \gamma_j f_{ih(t-j)}] = 1$. That is, in the linear combination (A3), the sum of the coefficients for the observations $y_{i,t}$, $i = 1, 2, \dots, s$, at time t is one. Therefore, at least one of the coefficients is greater than or equal to s^{-1} . Hence, $\text{Var}\{\sum_{j=0}^m \gamma_j \hat{\theta}_{t-j}(t)\} \geq s^{-2}v_{mm}$, and we conclude that $\sum_{(m)}$ is positive definite.

REFERENCES

- ADAM, A., and FULLER, W.A. (1992). Covariance estimators for the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 586-591.
- BAILAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.
- BELL, W.R., and HILLMER, S.C. (1990). The time series approach to estimation for periodic surveys. *Survey Methodology*, 16, 195-215.
- BINDER, D.A., and DICK, J.P. (1989). Modeling and estimation for repeated surveys. *Survey Methodology*, 15, 29-45.
- BREAU, P., and ERNST, L. (1983). Alternative estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.
- COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.
- FULLER, W.A. (1990). Analysis of repeated surveys. *Survey Methodology*, 16, 167-180.
- FULLER, W.A., ADAM, A., and YANSANEH, I.S. (1993). Estimators for longitudinal surveys. *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada, Ottawa, Canada, 309-324.
- JESSEN, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.
- KUMAR, S., and LEE, H. (1983). Evaluation of composite estimation for the Canadian Labour Force Survey. *Survey Methodology*, 9, 1-24.
- LENT, J., MILLER, S.M., and CANTWELL, P.J. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 130-139.
- ODELL, P.L., and LEWIS, T.O. (1971). Best linear recursive estimation. *Journal of the American Statistical Association*, 66, 893-896.

- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B* 12, 241-255.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.
- RAO, J.N.K., and GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SALLAS, W.M., and HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.
- SCOTT, A.J., SMITH, T.M.F, and JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *International Statistical Review*, 45, 13-28.
- SINGH, A. C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.
- TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.
- WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- YANSANEH, I.S. (1992). Least Squares Estimation for Repeated Surveys. Unpublished Ph.D. dissertation, Department of Statistics, Iowa State University, Ames, Iowa.
- YANSANEH, I.S. (1997). Recursive regression estimation in the presence of time-in-sample effects. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 164-169.

Estimation of Variance of General Regression Estimator: Higher Level Calibration Approach

SARJINDER SINGH, STEPHEN HORN and FRANK YU¹

ABSTRACT

In the present investigation, the problem of estimation of variance of the general linear regression estimator has been considered. It has been shown that the efficiency of the low level calibration approach adopted by Särndal (1996) is less than or equal to that of a class of estimators proposed by Deng and Wu (1987). A higher level calibration approach has also been suggested. The efficiency of higher level calibration approach is shown to improve on the original approach. Several estimators are shown to be the special cases of this proposed higher level calibration approach. An idea to find a non-negative estimate of variance of the GREG has been suggested. Results have been extended to a stratified random sampling design. An empirical study has also been carried out to study the performance of the proposed strategies. The well known statistical package, GES, developed at Statistics Canada can further be improved to obtain better estimates of variance of GREG using the proposed higher level calibration approach under certain circumstances discussed in this paper.

KEY WORDS: Calibration; Estimation of variance; Auxiliary information; Ratio and regression type estimators; Model assisted approach.

1. INTRODUCTION

The statisticians are often interested in the precision of survey estimates. The most commonly used estimator of population total/mean is the generalized linear regression (GREG) estimator. Let us consider the simplest case of the GREG where information on only one auxiliary variable is available. Consider a population $\Omega = \{1, 2, \dots, N\}$, from which a probability sample $s (s \subset \Omega)$ is drawn with a given sampling design, $p(\cdot)$. The inclusion probabilities $\pi_i = Pr(i \in s)$ and $\pi_{ij} \in Pr(i \text{ and } j \in s)$ are assumed to be strictly positive and known. Let y_i be the value of the variable of interest, y , for the i -th population element, with which also is associated an auxiliary variable x_i . For the elements, $i \in s$, we observe (y_i, x_i) . The population total of the auxiliary variable x , $X = \sum_{i=1}^N x_i$, is assumed to be accurately known. The objective is to estimate the population total $Y = \sum_{i=1}^N y_i$. Deville and Särndal (1992) used calibration on known population x -total to modify the basic sampling design weights, $d_i = 1/\pi_i$, that appear in the Horvitz-Thompson (1952) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{\pi_i} = \sum_{i=1}^n d_i y_i. \quad (1.1)$$

A new estimator,

$$\hat{Y}_{DS} = \sum_{i=1}^n w_i y_i \quad (1.2)$$

was proposed by Deville and Särndal (1992), with weights w_i as close as possible in an average sense for a given metric to the d_i , while respecting the calibration equation

$$\sum_{i=1}^n w_i x_i = X. \quad (1.3)$$

A simple case considered by Deville and Särndal (1992) is the minimization of chi-square type distance function given by

$$\sum_{i=1}^n \frac{(w_i - d_i)^2}{d_i q_i} \quad (1.4)$$

where q_i are suitably chosen weights. In most of the situations, the value of $q_i = 1$. The form of the estimator depends upon the choice of q_i . By minimizing (1.4) subject to calibration equation (1.3) we obtain weights

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i=1}^n d_i q_i x_i^2} \left(X - \sum_{i=1}^n d_i x_i \right). \quad (1.5)$$

Substitution of the value of w_i from (1.5) in (1.2) leads to the traditional regression estimator of total given by

$$\hat{Y}_{DS} = \sum_{i=1}^n d_i y_i + \frac{\sum_{i=1}^n d_i q_i x_i y_i}{\sum_{i=1}^n d_i q_i x_i^2} \left(X - \sum_{i=1}^n d_i x_i \right). \quad (1.6)$$

In this paper, the problem of estimation of variance of the regression estimator (1.6) has been considered at two different levels of calibration. The higher level calibration approach covers a greater variety of estimators than the low level calibration approach adopted by Särndal (1996).

¹ Sarjinder Singh, Research Officer, Stephen Horn, Senior Research Officer and Frank Yu, Director, Methodology Division, The Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT 2616, Australia.

Higher level calibration approach makes use of known total as well as known variance of the auxiliary character, whereas low level calibration utilizes only known total of auxiliary character.

The section 4 has been devoted to study the stratified sampling design. The original stratum weights are calibrated which results in combined regression and combined ratio estimators in stratified sampling. The estimators of variance of combined regression and combined ratio estimators proposed by Wu (1985) are shown to be the special cases of the low level calibration approach. The higher level calibration approach has been shown to apply to a broader variety of estimators.

2. ESTIMATOR OF VARIANCE OF THE GREG: THE LOW LEVEL CALIBRATION APPROACH

Following model assisted survey sampling approach of Särndal, Swensson and Wretman (1989, 1992), the Yates-Grundy (1953) form of estimator of variance of the estimator (1.6) is given by

$$\hat{V}_{YG}(\hat{Y}_{DS}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (w_i e_i - w_j e_j)^2 \quad (2.1)$$

where $D_{ij} = (\pi_i \pi_j - \pi_{ij}) / \pi_{ij}$, $i \neq j$ and $e_i = y_i - \hat{\beta} x_i$ have their usual meanings. This estimator can easily be written as

$$\begin{aligned} \hat{V}_{YG}(\hat{Y}_{DS}) &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i e_i - d_j e_j)^2 + \\ &\quad \hat{\psi}_1 \left(X - \sum_{i=1}^n d_i x_i \right) + \hat{\psi}_2 \left(X - \sum_{i=1}^n d_i x_i \right)^2 \end{aligned} \quad (2.2)$$

where

$$\begin{aligned} \hat{\psi}_1 &= \frac{1}{\sum_{i=1}^n d_i q_i x_i^2} \\ &\quad \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i e_i - d_j e_j) (d_i q_i x_i e_i - d_j q_j x_j e_j) \end{aligned} \quad (2.3)$$

and

$$\hat{\psi}_2 = \frac{1}{2 \left(\sum_{i=1}^n d_i q_i x_i^2 \right)^2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i q_i x_i e_i - d_j q_j x_j e_j)^2 \quad (2.4)$$

The estimator at (2.1) has been discussed by Särndal *et al.* (1989, 1992, 1996) on different occasions and covers a variety of estimators as discussed below:

For simplicity, let us consider simple random sampling and without replacement (SRSWOR) design *i.e.*, $\pi_i = \pi_j = n/N$ and $\pi_{ij} = n(n-1)/N(N-1)$. Then we have following cases:

Case 2.1: If $q_i = 1$, then (1.6) reduces to the usual regression estimator of total, \hat{Y}_{GREG} (say). Now if $w_i = d_i$ in (2.1), it reduces to

$$\hat{V}_{YG}(\hat{Y}_{GREG}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \quad (2.5)$$

where $f = n/N$ and $e_i = y_i - \hat{\beta} x_i$. Thus (2.5) denotes the usual estimator of variance of the regression estimator (1.6).

Case 2.2: If $q_i = 1/x_i$ then the estimator (1.6) reduces to the ratio estimator of total, \hat{Y}_{RATIO} (say). The estimator (2.1) reduces to an estimator of variance of the estimator \hat{Y}_{RATIO} , given by

$$\hat{V}_{YG}(\hat{Y}_{RATIO}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \left(\frac{X}{\hat{X}} \right)^2 \quad (2.6)$$

where

$$e_i = y_i - \left(\frac{\bar{y}}{\bar{x}} \right) x_i \text{ and } \hat{X} = \frac{N}{n} \sum_{i=1}^n x_i.$$

The estimator at (2.6) is a special case of a class of estimators of variance of the ratio estimator proposed by Wu (1982) as

$$\hat{V}_{YG}(\hat{Y}_W) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \left(\frac{X}{\hat{X}} \right)^g \quad (2.7)$$

for $g = 2$.

Case 2.3: If $q_i = 1$ and w_i is given by (1.5) then (2.2) and hence (2.1) becomes

$$\begin{aligned} \hat{V}_{YG}(\hat{Y}_{GREG}) &= \\ &\quad \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 + \hat{\psi}_1 (X - \hat{X}) + \hat{\psi}_2 (X - \hat{X})^2 \end{aligned} \quad (2.8)$$

where

$$\hat{\psi}_1 = \frac{(N-n)}{\left(\sum_{i=1}^n x_i^2 \right) n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (e_i - e_j) (x_i e_i - x_j e_j) \quad (2.9)$$

and

$$\hat{\psi}_2 = \frac{(N-n)}{2N(n-1) \left(\sum_{i=1}^n x_i^2 \right)^2} \sum_{i=1}^n \sum_{j=1}^n (x_i e_i - x_j e_j)^2. \quad (2.10)$$

Deng and Wu (1987) have defined a general class of estimators of the variance of the regression estimator as

$$\hat{V}_{YG}(\hat{Y}_{DW}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \left\{ \frac{X}{\hat{X}} \right\}^g \quad (2.11)$$

where $e_i = y_i - \hat{\beta}x_i$. The linear form of the class of estimators (2.11) takes the form as

$$\hat{V}_{YG}(\hat{Y}_{DW}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \left[1 + g \left(\frac{X}{\hat{X}} - 1 \right) + \frac{g(g-1)}{2} \left(\frac{X}{\hat{X}} - 1 \right)^2 + \dots \right] \quad (2.12)$$

which is again similar to (2.8). Thus the low level calibration approach considers estimators of variance of estimators of total *i.e.*, both ratio and regression methods of estimation. It is remarkable that there is no choice of q_i which reduces (1.6) to the product method of estimation considered by Cochran (1963). Hence the estimation of variance of product estimator has not been considered. To look at the efficiency of such estimators, we consider an analogue of the general class of estimators for estimating variance of GREG by following Srivastava (1971) as

$$\hat{V}_s(\hat{Y}_{GREG}) = \left(\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \right) H \left(\frac{X}{\hat{X}} \right) \quad (2.13)$$

where $H(\cdot)$ is a parametric function such that $H(1) = 1$ and satisfies certain regularity conditions. Following Srivastava (1971), it is easy to see that analogues of the general class of estimators (2.13) attain the minimum variance of the class of estimators proposed by Deng and Wu (1987) for regression estimator and Wu(1982) ratio estimator. We want to say here that if we will attach any function of the ratio X/\hat{X} to the usual estimator of variance given by

$$\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2,$$

the asymptotic variance of the resultant estimator remains the same. In other words, the efficiency of the estimators of variance of regression estimator (GREG) of total obtained through low level calibration remains less than or equal to the class of estimators proposed by Wu (1982) and Deng and Wu (1987). The weights w_i used to construct estimator of variance of GREG at (2.1) were obtained while estimating the population total and hence named as low level calibration weights for variance estimation. The next section is devoted to the higher level calibration approach where variance of auxiliary character is known. Several

new estimators are shown as special cases of the proposed higher level calibration approach.

3. IMPROVED ESTIMATOR OF VARIANCE OF THE GREG: THE HIGHER LEVEL CALIBRATION APPROACH

Here we apply the calibration approach to estimate the variance of GREG estimator at (1.6). The weights D_{ij} of Yates and Grundy (1953) for an estimator of variance given at (2.1) are calibrated such that the estimator of variance for the auxiliary variable has the exact variance. We consider an estimator of variance of GREG

$$\hat{V}_{SS}(\hat{Y}_{GREG}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \Omega_{ij} (w_i e_i - w_j e_j)^2 \quad (3.1)$$

where Ω_{ij} are the modified weights attached to the quadratic expression by Yates and Grundy (1953) form of estimator and are as close as possible in an average sense for a given measure to the D_{ij} with respect to the calibration equation

$$\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \Omega_{ij} (d_i x_i - d_j x_j)^2 = V_{YG}(\hat{X}_{HT}) \quad (3.2)$$

where

$$V_{YG}(\hat{X}_{HT}) = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N (\pi_i \pi_j - \pi_{ij}) (d_i x_i - d_j x_j)^2$$

denotes the known variance of the estimator of the auxiliary total $X (= \sum_{i=1}^N x_i)$ given by $\hat{X}_{HT} = \sum_{i=1}^n d_i x_i$. To compute the right hand side of (3.2) we need either information on every unit of the auxiliary character in the population, or only $V_{YG}(\hat{X}_{HT})$ obtained from a past survey or pilot survey. The examples of a situation where information on every unit of the auxiliary character is known are the establishment turnover recorded from census or administrative records or Business Register (BR) or Australian Taxation Office (ATO). Known variance of the auxiliary character has also been used by Das and Tripathi (1978), Singh and Srivastava (1980), Srivastava and Jhaji (1980, 1981), Isaki (1983), Singh and Singh (1988), Swain and Mishra (1992), Shah and Patel (1996) and Garcia and Cebrian (1996). Singh, Mangat and Mahajan (1995) have reviewed classes of estimators of unknown population parameters making use of the known variance of an auxiliary character. The idea of adjusting D_{ij} weights has also been discussed by Fuller (1970) through a regression type estimation procedure. For simplicity we restrict ourselves to the two dimensional Chi-Square (CS) type distance, D , between two $n \times n$ grids formed by the weights Ω_{ij} and D_{ij} for $i, j = 1, 2, \dots, n$, given by

$$D = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{(\Omega_{ij} - D_{ij})^2}{D_{ij} Q_{ij}} \quad (3.3)$$

In most of the situations $Q_{ij} = 1$ but other types of weights can also be used. We will show that the ratio type adjustment using known variance of auxiliary character is a special case for a particular choice of Q_{ij} . Minimization of (3.3) subject to (3.2) leads to modified optimal weights given by

$$\Omega_{ij} = D_{ij} + \frac{D_{ij} Q_{ij} (d_i x_i - d_j x_j)^2}{\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} Q_{ij} (d_i x_i - d_j x_j)^4} \left[V_{YG}(\hat{X}_{HT}) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i x_i - d_j x_j)^2 \right] \quad (3.4)$$

for the optimal choice of Lagrange Multiplier λ , given by

$$\lambda = \frac{V_{YG}(\hat{X}_{HT}) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i x_i - d_j x_j)^2}{\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} Q_{ij} (d_i x_i - d_j x_j)^4} \quad (3.5)$$

Its proof is given in the Appendix. Substitution of Ω_{ij} from (3.4) in (3.1) leads to the following regression type estimator,

$$\hat{V}_{SS}(\hat{Y}_{GREG}) = \hat{V}_{YG}(\hat{Y}_{DS}) + \hat{B}_1 \left[V_{YG}(\hat{X}_{HT}) - \hat{V}_{YG}(\hat{X}_{HT}) \right] \quad (3.6)$$

where

$$\hat{B}_1 = \frac{\sum_{i=1}^n \sum_{j=1}^n D_{ij} Q_{ij} (d_i x_i - d_j x_j)^2 (w_i e_i - w_j e_j)^2}{\sum_{i=1}^n \sum_{j=1}^n D_{ij} Q_{ij} (d_i x_i - d_j x_j)^4} = \frac{\hat{\mu}_{22}}{\hat{\mu}_{04}} \text{ (say)} \quad (3.7)$$

$\hat{V}_{YG}(\hat{X}_{HT}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i x_i - d_j x_j)^2$ and $\hat{V}_{YG}(\hat{Y}_{DS})$ is given in (2.1). Regression coefficient \hat{B}_1 makes use of the known total X of the auxiliary variable and hence can be treated as an improved estimator of regression coefficient by following Singh and Singh (1988). Under the higher level calibration approach, we have the following cases:

Case 3.1: Under SRSWOR sampling design if $q_i = x_i^{-1}$ and $Q_{ij} = (d_i x_i - d_j x_j)^{-2}$ are the weights attached at low level and higher level calibration approach, respectively, then the proposed strategy reduces to

$$\hat{V}_{SS}(\hat{Y}_{Ratio}) = \frac{N^2(1-f)}{n} \times \frac{1}{(n-1)} \sum_{i=1}^n e_i^2 \left(\frac{X}{\hat{X}} \right)^2 \left(\frac{S_x^2}{s_x^2} \right) \quad (3.8)$$

where $s_x^2 = (n-1)^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$ is an unbiased estimator of $S_x^2 = (N-1)^{-1} \sum_{i=1}^N (x_i - \bar{X})^2$.

Case 3.2: If $q_i = 1$ and $Q_{ij} = 1 \forall i \& j$, then we have

$$\hat{V}_{YG}(\hat{Y}_{GREG}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 + \hat{\psi}_1 (X - \hat{X}) + \hat{\psi}_2 (X - \hat{X})^2 + \hat{\psi}_3 (S_x^2 - s_x^2) \quad (3.9)$$

where $\hat{\psi}_1$ and $\hat{\psi}_2$ are given by (2.9) and (2.10), respectively, and

$$\hat{\psi}_3 = \frac{N^2(1-f)}{n \sum_{i=1}^n \sum_{j=1}^n (x_i - x_j)^4} \left[\sum_{i=1}^n \sum_{j=1}^n \left\{ (x_i - x_j)(e_i - e_j) + \frac{(X - \hat{X})(x_i - x_j)^2}{\sum_{i=1}^n x_i^2} \right\}^2 \right] \quad (3.10)$$

Without loss of generality, the estimators of variance of GREG given at (3.8) and (3.9) are neither members of a low level calibration approach nor of the class of estimators by Deng and Wu (1987). These estimators are members of the analogues of classes of estimators for estimating variance of GREG given by Srivastava and Jhajj (1981) as

$$\hat{V}_{SJ}(\hat{Y}_{GREG}) = \left(\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \right) H \left(\frac{X}{\hat{X}}, \frac{S_x^2}{s_x^2} \right) \quad (3.11)$$

where $H(.,.)$ is a parametric function such that $H(1, 1) = 1$ and which satisfies certain regularity conditions defined by them. Following Srivastava and Jhajj (1981) and Deng and Wu (1987), it is a class room exercise to see that the class of estimators at (3.11) remains better than the class of estimators defined at (2.11) and hence (2.13).

A difficult issue in using (3.1) is how to get non-negative estimates of variance using calibration. The simplest way is to optimize the CS distance function (3.3) subject to calibration constraint (3.2) along with the conditions $\Omega_{ij} \geq 0 \forall i, j = 1, 2, \dots, n$. While it is difficult to develop a solution to this problem theoretically, well known quadratic programming techniques can yield useful numerical results. Straightforward extension to using other distance functions, as discussed by Deville and Särndal (1992) for instance, to

the two dimensional problem due to the indeterminate nature of the D_{ij} weights is not possible. It is open to others to propose new distance functions which guarantee the non-negativity of the weights.

4. STRATIFIED SAMPLING DESIGN

Suppose the population consists of L strata with N_h units in the h -th stratum from which a simple random sample of size n_h is taken without replacement. The total population size $N = \sum_{h=1}^L N_h$ and sample size $n = \sum_{h=1}^L n_h$. Associated with the i -th unit of the h -th stratum there are two values y_{hi} and x_{hi} with $x_{hi} > 0$ being the covariate. For the h -th stratum, let $W_h = N_h/N$ be the stratum weights, $f_h = n_h/N_h$ the sample fraction, \bar{y}_h , \bar{x}_h , \bar{Y}_h , \bar{X}_h the y - and x - sample and population means respectively. Assume $\bar{X} = \sum_{h=1}^L W_h \bar{X}_h$ is known. The purpose is to estimate $\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h$, possibly by incorporating the covariate information x . The usual estimator of population mean \bar{Y} is given by

$$\bar{y}_{st} = \sum_{h=1}^L W_h \bar{y}_h. \quad (4.1)$$

We are considering a new estimator, given by

$$\bar{y}_{st}^* = \sum_{h=1}^L W_h^* \bar{y}_h \quad (4.2)$$

with new weights W_h^* . The new weights W_h^* are chosen such that chi-square type distance, given by

$$\sum_{h=1}^L \frac{(W_h^* - W_h)^2}{W_h q_h} \quad (4.3)$$

is minimum subject to the condition

$$\sum_{h=1}^L W_h^* \bar{x}_h = \bar{X}. \quad (4.4)$$

Minimization of (4.3) subject to calibration equation (4.4) leads to the combined regression type estimator given by

$$\bar{y}_{st}^* = \sum_{h=1}^L W_h \bar{y}_h + \frac{\sum_{h=1}^L W_h q_h \bar{x}_h \bar{y}_h}{\sum_{h=1}^L W_h q_h \bar{x}_h^2} \left[\bar{X} - \sum_{h=1}^L W_h \bar{x}_h \right] \quad (4.5)$$

for the optimum choice of weights given by

$$W_h^* = W_h + \frac{W_h q_h \bar{x}_h}{\sum_{h=1}^L W_h q_h \bar{x}_h^2} \left(\bar{X} - \sum_{h=1}^L W_h \bar{x}_h \right) \quad (4.6)$$

If $q_h = \bar{x}_h^{-1}$ then estimator (4.5) reduces to the well known combined ratio estimator in stratified sampling. The well known estimator of variance of combined regression estimator is given by

$$\hat{V}(\bar{y}_{st}^*) = \sum_{h=1}^L \frac{W_h^2 (1 - f_h)}{n_h} s_{eh}^2 \quad (4.7)$$

where

$$s_{eh}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} e_{hi}^2$$

is the h -th stratum sample variance and $\hat{e}_{hi} = y_{hi} - \bar{y}_h - b(x_{hi} - \bar{x}_h)$ and $b = \sum_{h=1}^L W_h q_h \bar{y}_h \bar{x}_h / \sum_{h=1}^L W_h q_h \bar{x}_h^2$ have their usual meaning. The lower level calibration approach yields an estimator of variance of the combined regression estimator as

$$\hat{V}_c(\bar{y}_{st}^*) = \sum_{h=1}^L \frac{D_h W_h^{*2}}{W_h^2} s_{eh}^2 \quad (4.8)$$

where

$$D_h = \frac{W_h^2 (1 - f_h)}{n_h}$$

and W_h^* is given by (4.6). If $q_h = \bar{x}_h^{-1}$ then (4.8) reduces to an estimator given by

$$\hat{V}(\bar{y}_{st}^*)_{\text{RATIO}} = \left(\frac{\bar{X}}{\bar{x}_{st}} \right)^2 \sum_{h=1}^L \frac{W_h^2 (1 - f_h)}{n_h} s_{eh}^2 \quad (4.9)$$

which is a special case of a class of estimators for estimating the variance of combined ratio estimator given by Wu (1985) as

$$\hat{V}(\bar{y}_{st}^*)_W = \left(\frac{\bar{X}}{\bar{x}_{st}} \right)^g \sum_{h=1}^L \frac{W_h^2 (1 - f_h)}{n_h} s_{eh}^2 \quad (4.10)$$

for $g = 2$. The properties of variance estimators of the combined ratio estimator are also studied by Saxena, Nigham and Shukla (1995). In higher level calibration, a new estimator is given by

$$\hat{V}_{st}(\hat{Y}_{\text{GREG}}) = \sum_{h=1}^L \frac{\Omega_h W_h^{*2}}{W_h^2} s_{eh}^2 \quad (4.11)$$

where Ω_h are suitably chosen weights such that Chi-Square distance function given by

$$\sum_{h=1}^L \frac{(\Omega_h - D_h)^2}{D_h Q_h} \quad (4.12)$$

is minimum subject to higher level calibration equation defined as

$$\sum_{h=1}^L \Omega_h s_{hx}^2 = V(\bar{x}_{St}) \quad (4.13)$$

where,

$$V(\bar{x}_{St}) = \sum_{h=1}^L W_h^2 \frac{(1-f_h)}{n_h} S_{hx}^2$$

is assumed to be known and $s_{hx}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ is an unbiased estimator of $S_{hx}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2$. This procedure leads to a new estimator for the variance of the combined regression estimator given by

$$\hat{V}(\hat{Y}_{St})_{CLR} = \hat{V}_{St}(\hat{Y}_{GREG}) + \hat{B}_{St} [V(\bar{x}_{St}) - \hat{V}(\bar{x}_{St})] \quad (4.14)$$

where

$$\hat{B}_{St} = \sum_{h=1}^L \frac{W_h^{*2} (1-f_h)}{n_h} Q_h s_{hx}^2 s_{eh}^2 / \sum_{h=1}^L \frac{W_h^2 (1-f_h)}{n_h} s_{hx}^4$$

denotes the combined improved estimator of regression coefficient in stratified sampling and

$$\hat{V}(\bar{x}_{St}) = \sum_{h=1}^L W_h^2 \frac{(1-f_h)}{n_h} s_{hx}^2$$

is an unbiased estimator of $V(\bar{x}_{St})$. If $q_h = 1/\bar{x}_h$ and $Q_h = 1/s_{hx}^2$, then estimator (4.14) reduces to a new estimator of variance of the combined ratio estimator given by

$$\hat{V}_{St}(\hat{Y}_{Ratio}) = \sum_{h=1}^L \frac{W_h^2 (1-f_h)}{n_h} s_{eh}^2 \left(\frac{\bar{X}}{\bar{x}_{St}} \right)^2 \left\{ \frac{V(\bar{x}_{St})}{\hat{V}(\bar{x}_{St})} \right\} \quad (4.15)$$

which is a ratio type estimator proposed by Wu (1985) for estimating variance of the combined ratio estimator but makes use of extra knowledge of the known variance of the auxiliary variable at the estimation stage. Several more new estimators can be constructed for new choices of weights q_h and Q_h .

5. A WIDER CLASS OF ESTIMATORS

If we define $u = X / \sum_{i=1}^n d_i x_i$ and $v = V(\hat{X}_{HT}) / \hat{V}(\hat{X}_{HT})$, then a wider class of estimators has been defined as

$$\hat{V}_{SS}(\hat{Y}_{GREG}) = \left\{ \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i e_i - d_j e_j)^2 \right\} H(u, v) \quad (5.1)$$

where $H(u, v)$ is a parametric function of u and v such that $H(1, 1) = 1$ and which satisfies certain regularity conditions. Then all estimators obtained from the following functions,

$$H(u, v) = u^\alpha v^\beta, \quad H(u, v) = \frac{1 + \alpha(u-1)}{1 + \beta(v-1)},$$

$$H(u, v) = 1 + \alpha(u-1) + \beta(v-1)$$

and $H(u, v) = \{1 + \alpha(u-1) + \beta(v-1)\}^{-1}$ are special cases of the higher level calibration approach, where α and β are unknown parameters involved in the function $H(u, v)$. Replacing these parameters with their respective consistent estimators in the class of estimators at (5.1) leads to the same asymptotic variance as shown by Srivastava and Jhajj (1983), Singh and Singh (1984) and Mahajan and Singh (1996). The extension of present investigation to two phase sampling following Hidiroglou and Särndal (1995) is in progress.

The next section has been devoted to studying the performance of the higher order calibration approach through simulation.

6. SIMULATION STUDY

Under the simulation study, we have considered comparisons of estimators of variance of ratio estimator as well as that of regression estimator. To avoid any kind of confusion, we have redefined the estimators considered for comparison as follows:

6.1 Ratio Estimator

We have compared the estimators of the variance of the ratio estimator, given by

$$\hat{V}_1(\hat{Y}_{Ratio}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \left(\frac{X}{\hat{X}} \right)^2 \quad (6.1.1)$$

with the estimator, given by

$$\hat{V}_2(\hat{Y}_{Ratio}) = \hat{V}_1(\hat{Y}_{Ratio}) \left(\frac{S_x^2}{s_x^2} \right). \quad (6.1.2)$$

6.2 Regression Estimator

We have also compared the estimators of the variance of the regression estimator, given by

$$\hat{V}_1(\hat{Y}_{GREG}) =$$

$$\frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 + \hat{\psi}_1 (X - \hat{X}) + \hat{\psi}_2 (X - \hat{X})^2 \quad (6.2.1)$$

with the estimator, given by

$$\hat{V}_2(\hat{Y}_{\text{GREG}}) = \hat{V}_1(\hat{Y}_{\text{GREG}}) + \hat{\Psi}_3(S_x^2 - s_x^2) \quad (6.2.2)$$

where $\hat{\Psi}_i, i = 1, 2, 3$ have the same meaning as defined earlier.

We have considered two types of populations viz. finite populations as well as infinite populations to cover almost all practical situations.

6.3 Finite Populations

In case of finite populations, we have taken a population consisting of $N = 20$ units from Horvitz and Thompson (1952). The study variable, y , is the number of households on i -th block and known auxiliary character, x , is the eye-estimated number of households on the i -th block. All possible samples of size $n = 5$ were selected by SRSWOR, which results in

$$\binom{N}{n} = 15,504$$

samples. From the k -th sample, the estimator

$$\hat{Y}_{\text{RATIO}}|_k = \hat{Y} \left(\frac{X}{\hat{X}} \right), \text{ with } \hat{Y} = \frac{N}{n} \sum_{i=1}^n y_i$$

was computed. Empirical mean squared error of this estimator was computed as

$$\text{MSE}(\hat{Y}_{\text{RATIO}}) = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} [\hat{Y}_{\text{RATIO}}|_k - Y]^2. \quad (6.3.1)$$

For the k -th sample, the ratio type estimators of variance

$$\hat{V}_h(\hat{Y}_{\text{RATIO}})|_k, h = 1, 2,$$

given by (6.1.1) and (6.1.2) respectively, for estimating the variance of the ratio estimator were also obtained. The bias in the h -th ratio type estimator of variance was computed as

$$B\{\hat{V}_h(\hat{Y}_{\text{RATIO}})\} = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} \hat{V}_h(\hat{Y}_{\text{RATIO}})|_k - \text{MSE}(\hat{Y}_{\text{RATIO}}) \quad (6.3.2)$$

and mean squared error was computed as

$$\text{MSE}\{\hat{V}_h(\hat{Y}_{\text{RATIO}})\} = \binom{N}{n}^{-1} \sum_{k=1}^{\binom{N}{n}} [\hat{V}_h(\hat{Y}_{\text{RATIO}})|_k - \text{MSE}(\hat{Y}_{\text{RATIO}})]^2. \quad (6.3.3)$$

The percent relative efficiency of the estimator $\hat{V}_2(\hat{Y}_{\text{RATIO}})$ with respect to $\hat{V}_1(\hat{Y}_{\text{RATIO}})$ was calculated as

RE =

$$\text{MSE}\{\hat{V}_1(\hat{Y}_{\text{RATIO}})\} \times 100 / \text{MSE}\{\hat{V}_2(\hat{Y}_{\text{RATIO}})\}. \quad (6.3.4)$$

The coverage by 95% confidence intervals

$$\text{CCI}[\hat{V}_h(\hat{Y}_{\text{RATIO}})]$$

for $h = 1, 2$ were calculated for h -th ratio type estimator of variance by counting the number of times the true population total, Y , falls between the limits defined as

$$\hat{Y}_{\text{RATIO}}|_k \mp t_{n-h-1}(\alpha) \sqrt{\hat{V}_h(\hat{Y}_{\text{RATIO}})|_k}. \quad (6.3.5)$$

These results were also obtained from all possible samples of size 6 and 7 and have been presented in Table 1.

The same process was repeated for the regression estimator

$$\hat{Y}_{\text{GREG}}|_k = \hat{Y} + \left(\sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 \right) (X - \hat{X})$$

of total obtained from (1.6) under a SRSWOR design. The biases, relative efficiency and CCI were obtained by using h -th estimator of variance of the regression estimator, $\hat{V}_h(\hat{Y}_{\text{GREG}})|_k$ for $h = 1, 2$, given by (6.2.1) and (6.2.2), respectively. The results obtained have been presented in Table 2. In addition, it was observed that for $n = 5$, 0.020% estimates of variance obtained from the estimator $\hat{V}_1(\hat{Y}_{\text{GREG}})$ and 0.022% estimates obtained from the estimator $\hat{V}_2(\hat{Y}_{\text{GREG}})$ were negative. Similar results were observed for more natural populations given by Cochran (1963) and Sukhatme and Sukhatme (1970). Over all, second order calibration estimators perform better than first order calibration in case of the finite populations.

In real life situations, the study variable and auxiliary variables may follow certain kinds of distributions like normal, beta or gamma *etc.* In order to see the performance of the proposed strategies under such circumstances, we generated artificial populations and considered the problem of estimation of finite population mean through simulation as follows.

Table 1
Comparison of $\hat{V}_2(\hat{Y}_{\text{RATIO}})$ with $\hat{V}_1(\hat{Y}_{\text{RATIO}})$ for finite populations

n	$B[\hat{V}_1(\hat{Y}_{\text{RATIO}})]$	$B[\hat{V}_2(\hat{Y}_{\text{RATIO}})]$	RE	$\text{CCI}[\hat{V}_1(\hat{Y}_{\text{RATIO}})]$	$\text{CCI}[\hat{V}_2(\hat{Y}_{\text{RATIO}})]$
5	-211.33	217.01	166.57	0.93	0.95
6	-141.92	102.00	115.06	0.91	0.92
7	-99.34	58.60	109.23	0.90	0.90

Table 2
Comparison of $\hat{V}_2(\hat{Y}_{\text{GREG}})$ and $\hat{V}_1(\hat{Y}_{\text{GREG}})$ for finite populations

n	$B[\hat{V}_1(\hat{Y}_{\text{GREG}})]$	$B[\hat{V}_2(\hat{Y}_{\text{GREG}})]$	RE	$\text{CCI}[\hat{V}_1(\hat{Y}_{\text{GREG}})]$	$\text{CCI}[\hat{V}_2(\hat{Y}_{\text{GREG}})]$
5	-328.49	-194.78	112.04	0.92	0.96
6	-223.92	-136.34	103.02	0.90	0.93
7	-157.88	-94.38	101.21	0.91	0.94

6.4 Infinite Populations

The size N of these populations is unknown. We generated n independent pairs of random numbers y_i^* and x_i^* (say), $i = 1, 2, \dots, n$, from a subroutine VNORM with $\text{PHI} = 0.7$, $\text{seed}(y) = 8987878$ and $\text{seed}(x) = 2348789$ following Bratley, Fox and Schrage (1983). For fixed $S_y^2 = 50$ and $S_x^2 = 50$, we generated transformed variables,

$$y_i = 3.0 + \sqrt{S_y^2(1 - \rho^2)} y_i^* + \rho S_y x_i^* \quad (6.4.1)$$

and

$$x_i = 4.0 + S_x x_i^* \quad (6.4.2)$$

for different values of the correlation coefficient ρ . For the k -th sample, the estimator

$$\hat{y}_{\text{RATIO}}|_k = \bar{y} \left(\frac{\bar{X}}{\bar{x}} \right), \text{ with } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ and } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

was computed. Empirical mean squared error of this estimator was computed as

$$\text{MSE}(\hat{y}_{\text{RATIO}}) = \frac{1}{15,000} \sum_{k=1}^{15,000} [\hat{y}_{\text{RATIO}}|_k - \bar{Y}]^2. \quad (6.4.3)$$

For the k -th sample, the ratio type estimators of variance

$$\hat{V}_h(\hat{y}_{\text{RATIO}})|_k, h = 1, 2,$$

obtained from (6.1.1) and (6.1.2) respectively, for estimating the variance of the ratio estimator of population mean were also derived. The bias in the h -th ratio type estimator of variance was computed as

$$B\{\hat{V}_h(\hat{y}_{\text{RATIO}})\} = \frac{1}{15,000} \sum_{k=1}^{15,000} \hat{V}_h(\hat{y}_{\text{RATIO}})|_k - \text{MSE}(\hat{y}_{\text{RATIO}}) \quad (6.4.4)$$

and mean squared error was computed as

$$\text{MSE}\{\hat{V}_h(\hat{y}_{\text{RATIO}})\} = \frac{1}{15,000} \sum_{k=1}^{15,000} [\hat{V}_h(\hat{y}_{\text{RATIO}})|_k - \text{MSE}(\hat{y}_{\text{RATIO}})]^2. \quad (6.4.5)$$

The percent relative efficiency of the estimator $\hat{V}_2(\hat{y}_{\text{RATIO}})$ with respect to $\hat{V}_1(\hat{y}_{\text{RATIO}})$ was calculated as

$$\text{RE} = \frac{\text{MSE}\{\hat{V}_1(\hat{y}_{\text{RATIO}})\} \times 100}{\text{MSE}\{\hat{V}_2(\hat{y}_{\text{RATIO}})\}} \quad (6.4.6)$$

The coverage by 95% confidence intervals

$$\text{CCI}[\hat{V}_h(\hat{y}_{\text{RATIO}})] \text{ for } h = 1, 2$$

was calculated for h -th ratio type estimator of variance by counting the number of times the true population mean, \bar{Y} , falls between the limits defined as

$$\hat{y}_{\text{RATIO}}|_k \mp 1.96 \sqrt{\hat{V}_h(\hat{y}_{\text{RATIO}})|_k}. \quad (6.4.7)$$

These results were obtained for samples of size $n = 60, 80$ and 100 for different values of correlation coefficient as presented in Table 3.

The same process was repeated for the regression estimator

$$\hat{\bar{y}}_{\text{GREG}} |_k = \bar{y} + \hat{\beta}(\bar{X} - \bar{x})$$

of mean obtained from (1.6) under a SRSWR design. The biases, relative efficiency and CCI were obtained by using h -th estimator of variance of the regression estimator,

$$\hat{V}_h(\hat{\bar{y}}_{\text{GREG}}) |_k \text{ for } h = 1, 2,$$

derived from (6.2.1) and (6.2.2), respectively. The results obtained have been presented in Table 4. We acknowledge that it is worth while studying the proposed strategy through simulation in more detail and its application in actual practice. The empirical study was carried out in FORTRAN-77 using a PENTIUM-120.

7. CONCLUSION

Higher level calibration approach can be used if variance of the auxiliary character is known in addition to the known total of that character. The statistical package GES developed by Statistics Canada can be modified to obtain better estimators of the variance of GREG, useful for surveys where information on variance of auxiliary characters is available or can be calculated.

ACKNOWLEDGEMENTS

The authors are heartily thankful to the Associate Editor and the two learned referees for fruitful and constructive comments to bring the original manuscript in the present form. They are also grateful to Dr. M.P. Singh for his kind suggestions. The opinions and results discussed in this paper are of authors and not necessarily of their institute(s).

Table 3
Comparison of $\hat{V}_2(\hat{\bar{y}}_{\text{RATIO}})$ with $\hat{V}_1(\hat{\bar{y}}_{\text{RATIO}})$ for infinite populations

n	ρ	$B[\hat{V}_1(\hat{\bar{y}}_{\text{RATIO}})]$	$B[\hat{V}_2(\hat{\bar{y}}_{\text{RATIO}})]$	RE	CCI $[\hat{V}_1(\hat{\bar{y}}_{\text{RATIO}})]$	CCI $[\hat{V}_2(\hat{\bar{y}}_{\text{RATIO}})]$
60	0.1	13.02	10.33	188.7	0.96	0.95
	0.3	8.07	6.35	192.6	0.97	0.95
	0.5	4.33	3.37	195.9	0.96	0.96
	0.7	1.77	1.37	197.9	0.97	0.97
	0.9	0.33	0.26	197.7	0.99	0.98
80	0.1	3.27	2.91	123.2	0.94	0.93
	0.3	2.06	1.84	123.0	0.94	0.94
	0.5	1.13	1.01	122.7	0.95	0.95
	0.7	0.47	0.42	122.0	0.97	0.96
	0.9	0.08	0.08	119.1	0.98	0.97
100	0.1	0.76	0.77	106.1	0.94	0.93
	0.3	0.49	0.49	105.8	0.94	0.94
	0.5	0.27	0.27	105.3	0.95	0.95
	0.7	0.12	0.12	104.4	0.96	0.95
	0.9	0.02	0.02	102.2	0.97	0.95

Table 4
Comparison of $\hat{V}_2(\hat{\bar{y}}_{\text{GREG}})$ with $\hat{V}_1(\hat{\bar{y}}_{\text{GREG}})$ for infinite populations

n	ρ	$B[\hat{V}_1(\hat{\bar{y}}_{\text{GREG}})]$	$B[\hat{V}_2(\hat{\bar{y}}_{\text{GREG}})]$	RE	CCI $[\hat{V}_1(\hat{\bar{y}}_{\text{GREG}})]$	CCI $[\hat{V}_2(\hat{\bar{y}}_{\text{GREG}})]$
60	0.1	10.12	8.42	177.6	0.98	0.95
	0.3	5.06	4.33	161.5	0.97	0.95
	0.5	3.32	2.36	152.5	0.95	0.96
	0.7	0.72	0.38	151.9	0.97	0.95
	0.9	0.13	0.10	147.7	0.99	0.97
80	0.1	1.23	1.11	153.9	0.96	0.95
	0.3	1.03	1.01	143.5	0.98	0.94
	0.5	0.13	0.11	132.8	0.97	0.95
	0.7	0.07	0.06	121.6	0.97	0.95
	0.9	0.02	0.03	117.1	0.96	0.96
100	0.1	0.65	0.57	136.1	0.95	0.94
	0.3	0.39	0.32	135.1	0.94	0.94
	0.5	0.13	0.13	129.6	0.95	0.95
	0.7	0.02	0.02	114.4	0.96	0.95
	0.9	0.01	0.01	112.2	0.97	0.96

APPENDIX

This appendix has been devoted to deriving the optimum value of Ω_{ij} as given in (3.4). The Lagrange's function is given by

$$L = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{(\Omega_{ij} - D_{ij})^2}{D_{ij} Q_{ij}} - 2\lambda \left[\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \Omega_{ij} (d_i x_i - d_j x_j)^2 - V_{YG}(\hat{X}_{HT}) \right]. \quad (A.1)$$

On differentiating (A.1) with respect to Ω_{ij} and equating to zero, we get

$$\Omega_{ij} = D_{ij} + \lambda D_{ij} Q_{ij} (d_i x_i - d_j x_j)^2. \quad (A.2)$$

On putting (A.2) in (3.2), we get

$$\lambda = \frac{V_{YG}(\hat{X}_{HT}) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i x_i - d_j x_j)^2}{\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} Q_{ij} (d_i x_i - d_j x_j)^4}. \quad (A.3)$$

On substituting (A.3) in (A.2), we get the optimum value of Ω_{ij} as given in (3.4).

REFERENCES

- BRATLEY, P., FOX, B.L., and SCHRAGE, L.E. (1983). *A Guide to Simulation*. New York: Springer-Verlag.
- COCHRAN, W.G. (1963). *Sampling Techniques*, (second edition). New York: John Wiley and Sons.
- DAS, A.K., and TRIPATHI, T.P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhyā*, 40(C), 139-148.
- DENG, L.Y., and WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.
- DEVILLE, J.-C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, 32, 209 - 226.
- GARCIA, M.R., and CEBRIAN, A.A. (1996). Repeated substitution method: The ratio estimator for the population variance. *Metrika*, 43, 101-105.
- HIDIROGLOU, M. A., and SÄRNDAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, Volume II, 873-878.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- ISAKI, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78(381), 117-123.
- MAHAJAN, P.K., and SINGH, S. (1996). On estimation of total in two stage sampling. *Journal of Statistical Research*, 30, 127-131.
- SÄRNDAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAXENA, S.K., NIGAM, A.K., and SHUKLA, N.D. (1995). Variance estimation for combined ratio estimator. *Sankhyā*, 57(B), 85-92.
- SHAH, D.N., and PATEL, P.A. (1996). Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling. *The Canadian Journal of Statistics*, 24(3), 373-384.
- SINGH, P., and SRIVASTAVA, S.K. (1980). Sampling scheme providing unbiased regression estimators. *Biometrika*, 67, 205-209.
- SINGH, R.K., and SINGH, G. (1984). A class of estimators with estimated optimum values in sample surveys. *Statistics & Probability Letters*, 2, 319-321.
- SINGH, S., and SINGH, S. (1988). Improved estimators of K and B in finite populations. *Journal of the Indian Society of Agricultural Statistics*, 121-126.
- SINGH, S., MANGAT, N.S., and MAHAJAN, P.K. (1995). General class of estimators. *Journal of the Indian Society of Agricultural Statistics*, 47(2), 129-133.
- SRIVASTAVA, S.K. (1971). A generalized estimator for the mean of finite population using multi-auxiliary information. *Journal of the American Statistical Association*, 66, 404-407.
- SRIVASTAVA, S.K., and JHAJJ, S.K. (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhyā* 42(C), 87-96.
- SRIVASTAVA, S.K., and JHAJJ, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, 68, 341-343.
- SRIVASTAVA, S.K., and JHAJJ, H.S. (1983). A class of estimators of estimators of the population mean using multi-auxiliary information. *Calcutta Statistical Association Bulletin* 32, 47-56.
- SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys With Applications*. Iowa: Iowa State University Press.
- SWAIN, A.K.P.C., and MISHRA, G. (1992). Unbiased estimators of finite population variance using auxiliary information. *Metron*, 201-215.
- WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.
- WU, C.F.J. (1985). Variance estimation for combined ratio and combined regression estimators. *Journal of the Royal Statistical Society*, 47(B), 147-154.
- YATES, F., and GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, 15(B), 253-261.

Logistic Generalized Regression Estimators

RISTO LEHTONEN and ARI VEIJANEN¹

ABSTRACT

In this paper we study the model-assisted estimation of class frequencies of a discrete response variable by a new survey estimation method, which is closely related to generalized regression estimation. In generalized regression estimation the available auxiliary data are incorporated in the estimation procedure by a linear model fit. Instead of using a linear model for the class indicators, we describe the joint distribution of the class indicators by a multinomial logistic model. Logistic generalized regression estimators are introduced for class frequencies in a population and domains. Monte Carlo experiments were carried out for simulated data and for real data taken from the Labour Force Survey conducted monthly by Statistics Finland. The logistic generalized regression estimation yielded better results than the ordinary regression estimation for small domains and particularly for small class frequencies.

KEY WORDS: Auxiliary information; Class frequencies; Generalized linear models; Labour force survey; Model-assisted estimation; Regression estimators.

1. INTRODUCTION

Consider the estimation of class frequencies of a discrete response variable in a sample survey. The number of individuals in a class equals the class indicator's sum over the population, the total of the indicator. Therefore, the problem can be solved by methods designed for the estimation of population totals. To improve the accuracy of the estimation, a survey statistician often makes use of the available auxiliary data. If the expectation of the response variable can be assumed to depend linearly on the auxiliary variables as can be the case for continuous response variables, it is advisable to use the generalized regression estimator (Särndal, Swensson and Wretman 1992; Estevao, Hidioglou and Särndal 1995). Generalized regression estimation can improve the efficiency and reduce the bias due to unit nonresponse if the auxiliary variables correlate strongly with the response variable.

From a modeler's perspective, a linear model is quite restrictive and might not be the best choice for binary response variables, such as employment status of a person (employed, unemployed), or more generally for discrete response variables, such as a person's status in the labour market (employed, unemployed, not in labour force). For such variables we introduce a class of logistic generalized regression estimators based on a multinomial logistic model describing the joint distribution of the class indicators. The motivation for the selection of this specific model type thus is similar to that used in the context of generalized linear models (McCullagh and Nelder 1989).

The parameters of the logistic model are here estimated by maximizing a sample-based weighted loglikelihood, the Horvitz-Thompson estimator of the population loglikelihood function (Godambe and Thompson 1986; Nordberg

1989; Skinner, Holt and Smith 1989; Särndal *et al.* 1992, p. 517).

As an application, we consider the estimation of the unemployment rate in the Labour Force Survey conducted monthly by Statistics Finland. Administrative records indicating whether a person is registered jobseeker in local employment office are available as register-based auxiliary data, and these records were merged with the survey data on individual basis using personal identification numbers which are unique in both data sources. The corresponding auxiliary variable correlates strongly with the survey measurement on person's unemployment. Thus, improvement in efficiency and reduction of bias can be expected by making use of these administrative data in the estimation procedure. Additional auxiliary data (sex, age, regional data) were gathered from the Population Register. Also these auxiliary data were merged with the survey data on individual basis.

The properties of the generalized regression estimators were studied by Monte Carlo simulation methods where SRSWOR samples were repeatedly drawn from a population constructed from the Labour Force Survey data. We use incomplete poststratification or raking based on a main effects ANOVA model. The experiments indicate that the logistic formulation yields better results than the linear formulation for small domains. We obtained good results also when there was only one continuous auxiliary variable.

This paper is organized as follows. Section 2 defines the multinomial logistic model and basic concepts used. In Section 3 we introduce generalized regression estimators of class frequencies in a population and domains, and discuss the estimation of the model parameters by weighted loglikelihood. Variance estimators are presented. Monte Carlo experiments are discussed in Section 4. Conclusions are drawn in Section 5.

¹ Risto Lehtonen and Ari Veijanen, Statistics Finland, P.O. Box 5A, FIN-00022 Statistics Finland, Finland.

2. MODEL

Consider discrete m -valued random variables Y_k associated with N elements k in a finite population U . We observe their realized values y_k only in a sample $s \subset U$ of size n . Our goal is to estimate the frequency distribution of the y_k 's in the population; in classification problems, we estimate the class proportions. Suppose we know the vector of auxiliary variables \mathbf{x}_k for every element in the population. We impose a multinomial logistic model

$$P\{Y_k=i\} = \frac{\exp\{\mathbf{x}_k' \boldsymbol{\beta}_i\}}{\sum_{r=1}^m \exp\{\mathbf{x}_k' \boldsymbol{\beta}_r\}} \quad (i = 1, 2, \dots, m) \quad (1)$$

and assume that the Y_k 's are conditionally independent given the \mathbf{x}_k 's. In the binary case, this is the model used in logistic regression. The parameter vector $\boldsymbol{\beta}$ is composed of vectors $\boldsymbol{\beta}_i (i = 1, 2, \dots, m)$ with components $\beta_{ij} (j = 1, 2, \dots, q)$. The parameters are assumed identifiable, that is, no two parameter values yield identical probabilities (1) for every k . This implies that the auxiliary variables $x_{kj} (j = 1, 2, \dots, q)$ are linearly independent. To avoid identifiability problems, we set $\boldsymbol{\beta}_1 = 0$. It is straightforward to generalize (1) so that different auxiliary variables can be assigned for the m classes (Lehtonen and Veijanen 1998).

The sampling design specifies the inclusion probabilities of population elements. The k -th element is drawn with inclusion probability π_k and elements k and p are simultaneously in the sample s with probability $\pi_{kp} > 0$ ($\pi_{kk} = \pi_k$). As usual, the sample membership indicators $I_k = I\{k \in s\}$ are assumed conditionally independent of the Y_k 's given the \mathbf{x}_k 's, but the inclusion probabilities may correlate with the auxiliary variables.

Under unit nonresponse, if element k responds with probability θ_k independently of the I_p 's and Y_p 's ($p \in U$), then we substitute $\pi_k \theta_k$ for π_k . Correspondingly, π_{kp} is replaced by $\pi_{kp} \theta_k \theta_p$ when the elements respond independently of each other.

3. LOGISTIC GENERALIZED REGRESSION ESTIMATION

3.1 Definition of LGREG

To estimate the frequency distribution of the y_k 's, we define class indicators $Z_{ki} = I\{Y_k = i\}$ with realizations z_{ki} and estimate the totals $t_i = \sum_{k \in U} z_{ki}$. The Horvitz-Thompson (HT) estimator of t_i is $\hat{t}_i^{\text{HT}} = \sum_{k \in s} a_k z_{ki}$, where the sampling weights are $a_k = 1/\pi_k$. Generalized regression estimation (GREG) is assisted by a regression model $Z_{ki} = \mathbf{x}_k' \boldsymbol{\beta}_i^G + \varepsilon_{ki}$ with $\text{Var}(\varepsilon_{ki}) = \sigma_{ki}^2$ (Särndal *et al.* 1992; Estevao *et al.* 1995). The parameter $\boldsymbol{\beta}_i^G$ is estimated by

$$\hat{\boldsymbol{\beta}}_i^G = \left(\sum_{k \in s} a_k \frac{\mathbf{x}_k \mathbf{x}_k'}{\sigma_{ki}^2} \right)^{-1} \left(\sum_{k \in s} a_k \frac{\mathbf{x}_k z_{ki}}{\sigma_{ki}^2} \right) \quad (i = 1, 2, \dots, m) \quad (2)$$

and the fitted values $\hat{z}_{ki} = \mathbf{x}_k' \hat{\boldsymbol{\beta}}_i^G$ are incorporated in the GREG estimator

$$\hat{t}_i^G = \sum_{k \in U} \hat{z}_{ki} + \sum_{k \in s} a_k (z_{ki} - \hat{z}_{ki}) \quad (i = 1, 2, \dots, m). \quad (3)$$

The selection of a linear model for a GREG estimator (3) is fully justified for a continuous response variable. For binary measurements Z_{ki} , a linear model might be unrealistic. Ordinarily, we would prefer a logistic model to a linear one. In the logistic formulation, the predicted value always lies in $[0, 1]$, whereas in the linear formulation, the predicted value can exceed these natural limits. If the probability of $Z_{ki} = 1$ is close to 0 or 1, then the two models yield different results. Moreover, when there are $m > 2$ classes, it appears sensible to describe the joint distribution of the Z_{ki} 's ($i = 1, 2, \dots, m$) by the multinomial logistic model (1). To apply the model (1) in generalized regression estimation, we estimate the expectations $\mu_{ki} = E(Z_{ki} | \mathbf{x}_k; \boldsymbol{\beta}) = P\{Y_k = i | \mathbf{x}_k; \boldsymbol{\beta}\}$ by

$$\hat{\mu}_{ki} = P\{Y_k = i | \mathbf{x}_k; \hat{\boldsymbol{\beta}}\} = \frac{\exp\{\mathbf{x}_k' \hat{\boldsymbol{\beta}}_i\}}{1 + \sum_{r=2}^m \exp\{\mathbf{x}_k' \hat{\boldsymbol{\beta}}_r\}},$$

which depend nonlinearly on the auxiliary variables. We define a logistic generalized regression (LGREG) estimator by

$$\hat{t}_i = \sum_{k \in U} \hat{\mu}_{ki} + \sum_{k \in s} a_k (z_{ki} - \hat{\mu}_{ki}) \quad (i = 1, 2, \dots, m). \quad (4)$$

The GREG and LGREG estimators (3) and (4) include a sum of predicted values over the population. However, it is not actually necessary to have information about the \mathbf{x}_k 's for every element in the population U . In GREG (3), it is enough to know the auxiliary totals $\sum_{k \in U} \mathbf{x}_k$, because (3) can also be expressed in the form $\hat{t}_i^G = \hat{t}_i^{\text{HT}} + (\sum_{k \in U} \mathbf{x}_k - \sum_{k \in s} a_k \mathbf{x}_k)' \hat{\boldsymbol{\beta}}_i^G$. For the special case of complete poststratification, the information required in LGREG is similar to that needed in GREG. For other cases, such as incomplete poststratification, we cannot compute $\sum_{k \in U} \hat{\mu}_{ki}$ in (4) without knowing the frequency of each value of \mathbf{x}_k in the population. For example, if we have two discrete auxiliary variables, then in GREG we need the marginal frequencies, but in LGREG we need the cell frequencies.

In addition to estimates for the whole population, estimates are usually calculated for subpopulations. The population U is partitioned into domains $U_{(d)} \subset U$ of size

$N_{(d)}$. The set s of respondents is composed of corresponding subsets $s_{(d)} = s \cap U_{(d)}$ with $n_{(d)}$ elements. As in GREG estimation (Särndal *et al.* 1992), we apply LGREG estimator

$$\hat{t}_{(d)i} = \sum_{k \in U_{(d)}} \hat{\mu}_{ki} + \sum_{k \in s_{(d)}} a_k (z_{ki} - \hat{\mu}_{ki}). \quad (5)$$

These estimators are additive: $\sum_i \hat{t}_{(d)i} = N_{(d)}$. If we combine two nonoverlapping domains d_1 and d_2 , the LGREG estimate for $d = d_1 \cup d_2$ is $\hat{t}_{(d)i} = \hat{t}_{(d_1)i} + \hat{t}_{(d_2)i}$. Hence, $\sum_d \hat{t}_{(d)i} = \hat{t}_i$ for nonoverlapping domains and $\sum_i \hat{t}_i = N$.

In generalized regression estimation, an estimate (3) or (4) can be negative, when negative residuals coincide with large values of a_k . Negative GREG estimates become more common, as the number of auxiliary variables increases (Chambers 1996). In LGREG estimation, in contrast, this is not so, because $\hat{\mu}_{ki}$ is bounded by the model formulation. In our experiments, LGREG estimates were negative only for small domains in certain cases. In many cases, LGREG estimate equals the sum of estimated expectations and then it is always positive (see Section 3.2).

If the model (1) includes an auxiliary indicator variable, its total over the population is exactly estimated by LGREG. This calibration property is desirable in many applications.

3.2 ML Estimation by π -Weighted Loglikelihood

We estimate the parameter β in the model (1) by maximizing a π -weighted loglikelihood

$$L_s(\beta_2, \dots, \beta_m) = \sum_{k \in s} \pi_k^{-1} \left\{ I\{Y_k = 1\} \log \left(1 - \sum_{i=2}^m \mu_{ki} \right) + \sum_{i=2}^m I\{Y_k = i\} \log \mu_{ki} \right\}$$

(Godambe and Thompson 1986; Nordberg 1989; Särndal *et al.* 1992, p. 517). In general, we maximize the likelihood function numerically by appropriate numerical methods such as a Newton-Raphson algorithm.

It can be shown that for complete poststratification, the fitted values \hat{z}_{ki} in GREG are equal to the estimates $\hat{\mu}_{ki}$ in LGREG. Thus, when there are no missing cells in complete poststratification, the GREG and LGREG estimators are identical (Lehtonen and Veijanen 1998). This does not hold for other models such as incomplete poststratification.

The LGREG estimator (4) has two parts: a sum of estimated expectations over the population and an adjustment term $\sum_{k \in s} a_k (z_{ki} - \hat{\mu}_{ki})$. It can be shown that if the model contains an intercept, the adjustment term vanishes and the frequency t_i is estimated by $\sum_{k \in U} \hat{\mu}_{ki}$ (Lehtonen and Veijanen 1998).

In our experiments, we apply a ratio estimator $\hat{R} = \hat{t}_i / (\hat{t}_i + \hat{t}_j)$. Its variance is estimated by Taylor linearization techniques (Särndal *et al.* 1992, p. 179):

$$\hat{V}(\hat{R}) = \frac{1}{(\hat{t}_i + \hat{t}_j)^2} \left[(1 - \hat{R})^2 \hat{C}_{ii} + 2\hat{R}(\hat{R} - 1) \hat{C}_{ij} + \hat{R}^2 \hat{C}_{jj} \right], \quad (6)$$

where C_{ij} , the covariance of \hat{t}_i and \hat{t}_j , is estimated by

$$\hat{C}_{ij} = \sum_{k, p \in s} \frac{\Delta_{kp} e_{ki} e_{pj}}{\pi_{kp} \pi_k \pi_p}. \quad (7)$$

In (7), $e_{ki} = z_{ki} - \hat{\mu}_{ki}$ and $\Delta_{kp} = \text{Cov}(I_k, I_p) = \pi_{kp} - \pi_k \pi_p$. Similar derivations hold for the corresponding domain estimators.

4. EXPERIMENTS

4.1 Details of Simulation Studies

In all the simulation experiments, $K = 1,000$ samples were drawn from a population with simple random sampling without replacement (SRSWOR). Monte Carlo means and standard errors of the estimates were calculated from the simulated samples. The design effect for an estimator $\hat{t}_{(d)i}$ was calculated as a ratio of estimated variances: $\text{Deff}(\hat{t}_{(d)i}) = \hat{V}_{mc}(\hat{t}_{(d)i}) / \hat{V}_{mc}(\hat{t}_{(d)i}^{\text{HT}})$, where $\hat{V}_{mc}(\hat{t}_{(d)i}^{\text{HT}})$ denotes the Monte Carlo variance estimate of the HT estimator (Lehtonen and Pahkinen 1996). We measured the overall accuracy of domain estimates by the mean absolute relative domain error over D domains and K samples s_j :

$$\text{MARDE}(i) = \frac{1}{D} \sum_{p=1}^D \frac{1}{K} \sum_{j=1}^K \frac{100 \left| \hat{t}_{(d_p)i}(s_j) - t_{(d_p)i} \right|}{t_{(d_p)i}}.$$

In the GREG estimates (2), the variance was a constant $\sigma_{ki}^2 = \sigma^2$, which cancelled out. For LGREG, domain frequencies were estimated by (5) and variances by (7). For GREG and HT, see Särndal *et al.* (1992, p. 401). Confidence intervals for the frequencies were computed as if the class indicators were independent. At the nominal significance level of 95%, an acceptable coverage rate lies in [93.65%, 96.35%] for $K = 1,000$ simulated samples.

4.2 An Experiment With Simulated Data

To compare LGREG with GREG, we simulated a data set, in which the auxiliary variable X was a continuous random variable uniformly distributed in $(-3, 3)$. The variable of interest, Y , representing three classes followed distribution (1) specified by $x'_k \beta_1 = 0$, $x'_k \beta_2 = 3X_k - 1$, and $x'_k \beta_3 = -2X_k$ for $N = 10,000$ elements ($k = 1, 2, \dots, N$). A

thousand samples of size $n = 1,000$ were independently drawn with SRSWOR. X_k and X_k^2 were used as auxiliary variables. All the estimators appeared unbiased (Table 1). The variance estimates had empirical bias smaller than 3% and standard deviation smaller than 5%.

Table 1

The design effects (Deff) for class frequency estimators and the empirical coverage rates (CR) (%) of 95% confidence intervals for classes $i = 1, 2, 3$

Method	Deff			CR		
	\hat{t}_1	\hat{t}_2	\hat{t}_3	\hat{t}_1	\hat{t}_2	\hat{t}_3
HT	1	1	1	95.2	95.3	94.7
GREG	0.93	0.55	0.57	95.0	94.3	95.6
LGREG	0.89	0.45	0.50	94.9	93.7	95.3

The best results were obtained by LGREG, probably due to the fact that the proportional frequencies of classes varied greatly over the range of the auxiliary variable. The probability of each class was such a function of the continuous auxiliary variable that a linear regression model did not fit the data well.

4.3 An Experiment With the Finnish Labour Force Survey Data

4.3.1 Constructed Population

We studied the estimation of the unemployment rate using the Finnish Labour Force Survey (LFS) data of three consecutive months of the year 1994. The constructed population consisted of 33,329 individuals. From the Population Register we obtained, for each population member, age class (15-24, 25-34, 35-44, 45-54, and 55-64 years), sex and region (three areas). A jobseeker indicator was obtained from the register maintained by Ministry of Labour showing which individuals were registered as unemployed jobseekers. The time lag in this administrative data source is about two weeks. It can thus be expected that the proportion of persons with changes in the actual labour market status is small within this short time interval. It should be noticed that the register-based jobseeker status is defined differently from the employment status measured in the Labour Force Survey. The survey measurement is based on a standard International Labour Office (ILO) definition. All these auxiliary data were merged with the survey data on individual basis.

The nonresponse rate varied by jobseeker status so that among registered jobseekers the rate was 11.4% whereas for the others the rate was 7.6%. The probability of nonresponse was modeled by a logistic ANOVA model and the ML estimates of nonresponse rates (ranging from 2.9% to 22.8%) were used as a nonresponse model in simulations.

For simulation experiments, we constructed an artificial population consisting of $N = 30,835$ persons. Employment status was defined by three classes: "employed", "unemployed", and "not in labour force" with population frequencies $t_1 = 17,373$, $t_2 = 4,433$, and $t_3 = 9,029$, respectively. The unemployment rate was defined by $R = t_2 / (t_1 + t_2) = 20.33\%$. As domains we used the cells in the crosstabulation of age classes, sex, and the register-based unemployment status.

From the artificial population, $K = 1,000$ independent random samples of size $n = 1,000$ persons were drawn with simple random sampling without replacement. In each sample, nonresponse was simulated by the nonresponse model fitted to the original population. The response probabilities were then estimated from each sample by logistic regression with the same ANOVA model as in the nonresponse model. We multiplied each probability π_k by the estimated response probability.

Three models were used to compare LGREG with GREG. The components of x_k were dummies corresponding to age (5 classes), sex, region (3 areas) and jobseeker status. In incomplete poststratification, or raking, a main effects ANOVA model was based on classified auxiliary variables. We compared models with and without the jobseeker indicator. The third model also included a fourth-order polynomial of age.

4.3.2 Results

Incorporating no auxiliary information, HT estimators had usually larger variance than the generalized regression estimators (Table 2). Both generalized regression estimators based on a raking model with age, sex, and region yielded some improvement over the HT estimates. Much better results were obtained by models including the jobseeker indicator, which correlates more strongly ($r = 0.83$) with the ILO unemployment indicator than the other auxiliary variables. Thus these auxiliary data improve the efficiency of estimation (cf. Djerf 1997).

Table 2

Properties of unemployment rate estimates ($\hat{R}(\%)$) for the raking model (R) and the model including age polynomial (P), with (E) or without (N) the jobseeker indicator. SD denotes the standard deviation and CR (%) denotes the coverage rate of 95% confidence intervals

Model	Method	\hat{R}	Bias	SD	Deff	CR	MARDE
	HT	20.32	-0.0081	1.461	1	95.7	35.28
RN	GREG	20.30	-0.0262	1.454	0.995	95.3	46.03
RN	LGREG	20.31	-0.0229	1.454	0.995	95.3	45.93
RE	GREG	20.30	-0.0244	0.895	0.612	96.0	35.74
RE	LGREG	20.29	-0.0419	0.901	0.617	94.8	34.80
PE	GREG	20.30	-0.0259	0.887	0.607	95.6	35.41
PE	LGREG	20.29	-0.0421	0.896	0.613	95.1	34.76

Table 3

Mean absolute relative domain errors (MARDE) and mean coverage rates (CR) (%) of 95% confidence intervals for estimated class frequencies in domains with true frequency $t_{(d)i}$ ($i = 1, 2, 3$) (a) smaller than 100, and (b) at least 100. The model included the age polynomial

	Method	MARDE			CR		
		$\hat{t}_{(d)1}$	$\hat{t}_{(d)2}$	$\hat{t}_{(d)3}$	$\hat{t}_{(d)1}$	$\hat{t}_{(d)2}$	$\hat{t}_{(d)3}$
(a)	GREG	96.92	67.36	121.95	88.2	77.8	84.6
	LGREG	80.28	67.20	104.05	83.9	76.5	51.7
(b)	GREG	6.95	12.31	14.35	94.1	85.9	93.7
	LGREG	6.88	12.34	14.29	93.9	85.4	93.3

The differences between GREG and LGREG were small at the population level (Table 2). LGREG was never inferior to GREG. Domain totals, especially in small domains, were more accurately estimated by LGREG than by GREG (Table 3). When the model included the age as a continuous auxiliary variable, the standard deviation of the unemployment rate estimate was smaller for LGREG than for GREG in 19 of 20 domains. Unfortunately, the confidence intervals obtained by LGREG were often too narrow due to small variance estimates (Table 3).

5. SUMMARY

We introduce a new approach to the model-assisted estimation of population class frequencies of a discrete response variable in survey sampling. Our logistic generalized regression estimation (LGREG) is based on a multinomial logistic model, which might be more realistic for class indicators than the linear model normally used in generalized regression estimation (GREG). LGREG and GREG estimators yield identical results for complete poststratification, but differ for other models such as raking. As compared with GREG, LGREG usually requires more auxiliary information, not only the auxiliary totals. Nevertheless, LGREG appears preferable to GREG when the class probabilities vary greatly over the range of continuous auxiliary variables and when we need estimates for small

domains, particularly in the presence of small class frequencies.

ACKNOWLEDGEMENTS

We are grateful to Prof. Carl-Erik Särndal, University of Montreal, for his comments on the previous version of the manuscript. Detailed comments given by an Associate Editor and two referees were very helpful. Thanks are also due to Mr. Timo Koskimäki, Statistics Finland, for providing us with the Labour Force Survey data and to Mr. Kari Djerf for helpful comments.

REFERENCES

- CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.
- DJERF, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*, 13, 29-39.
- ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- GODAMBE, V.P., and THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *International Statistical Review*, 54, 127-138.
- LEHTONEN, R., and PAHKINEN, E.J. (1996). *Practical Methods for Design and Analysis of Complex Surveys*. Revised Edition. Chichester: John Wiley & Sons.
- LEHTONEN, R., and VEIJANEN, A. (1998). On Multinomial Logistic Generalized Regression Estimators. Jyväskylä. Preprints from the Department of Statistics, University of Jyväskylä, 22.
- McCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models*. Second Edition. London: Chapman and Hall.
- NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, 5, 223-239.
- SKINNER, C.J., HOLT, D., and SMITH, T.M.F. (Eds) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Confidence Intervals for Domain Parameters When the Domain Sample Size is Random

ROBERT J. CASADY, ALAN H. DORFMAN and SUOJIN WANG¹

ABSTRACT

Let A be a population domain of interest and assume that the elements of A cannot be identified on the sampling frame and the number of elements in A is not known. Further assume that a sample of fixed size (say n) is selected from the entire frame and the resulting domain sample size (say n_A) is random. The problem addressed is the construction of a confidence interval for a domain parameter such as the domain aggregate $T_A = \sum_{i \in A} x_i$. The usual approach to this problem is to redefine x_i , by setting $x_i = 0$ if $i \notin A$. Thus, the construction of a confidence interval for the domain total is recast as the construction of a confidence interval for a population total which can be addressed (at least asymptotically in n) by normal theory. As an alternative, we condition on n_A and construct confidence intervals which have approximately nominal coverage under certain assumptions regarding the domain population. We evaluate the new approach empirically using artificial populations and data from the Bureau of Labor Statistics (BLS) Occupational Compensation Survey.

KEY WORDS: Bayes method; Conditioning; Establishment surveys; Simple random sampling; Stratification; Survey methods.

1. INTRODUCTION

In sampling from a finite population, we often are interested in the estimation of totals, means, or other quantities, for parts of that population, usually referred to as domains. Such domains are not explicitly listed in the frame, the number of items that will occur in the survey is not known in advance, and often enough, we do not even know the number of their elements in the population. For example, we might sample schoolchildren for certain medical problems, and then wish to know the mean blood pressure of those children who are underweight. The class of underweight children would constitute a domain. The only information we have as to whether or not a child is underweight is likely to be among the sampled children; if so, then this would be a case where the domain is not explicitly listed on the frame.

An essential part of the inference process is the estimation of the precision of our estimators; this is typically given by an estimated standard deviation, coefficient of variation, or confidence interval. The notion of a valid confidence interval underlies whatever measure of precision we use. All confidence intervals have, by construction, a stated "nominal" confidence level. A valid confidence interval is a confidence interval with actual coverage matching the nominal coverage. The actual coverage may be determined theoretically or by empirical work mimicking the practical circumstances in which the confidence interval would be used. If a standard deviation is not such as to give rise to a valid confidence interval, then the standard deviation needs to be regarded as misleading.

In the case of estimates for domains, confidence intervals constructed along traditional lines can lead to serious under-coverage, a fact not always appreciated in the literature. We refer to this as the domain problem. The present paper addresses this problem by a somewhat complex methodology involving Bayesian ideas, which, however, leads to a rather simple practical solution, improving on current methodology. The main change in method lies in replacing the standard normal statistic used in the construction of confidence intervals, with a Student's t -statistic having degrees of freedom that depend on the number and configuration of the domain items in the sample.

We shall focus on domain totals and domain means for the two common cases of simple random sampling and stratified random sampling. In the case of simple random sampling, it turns out that standard methods are satisfactory for the mean; however, for the total, coverage can be lower than nominal but not usually worrisome. For stratified random sampling, confidence intervals for both the mean and the total pose serious difficulties with regard to coverage level. In this case, the new methodology is augmented by use of a well known approximation due to Satterthwaite (1946). Alternate approaches to ours, also using this approximation, may be found in Johnson and Rust (1993) and Kott (1994).

An outline of the paper is as follows: In Section 2, to introduce ideas, we consider the case of the total in simple random sampling, using it to illustrate the standard approach for domain estimation, the coverage problem to which this gives rise, and the approach here taken to rectify the difficulty. Section 3 describes the extension to stratified random sampling. Section 4 states our conclusions.

¹ Robert J. Casady and Alan H. Dorfman, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001, U.S.A.; Suojin Wang, Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

2. THE CASE OF SIMPLE RANDOM SAMPLING

2.1 Standard Method

The standard approach to domain estimation is well described in Särndal, Swensson, and Wretman (1992); Sections 3.3, 5.8, and Chapter 10) (henceforth SSW). Their approach is general. Here we paraphrase it for the case of simple random sampling, and, by mild extension, for stratified random sampling as well, and focus on the domain total.

Let x_i be the value of the characteristic of interest for the i -th ($i = 1, 2, \dots, N$) element of the population and let A be a domain of interest. We shall consider only the case where the elements of A cannot be identified on the frame and the number N_A of elements in A is not known; the case where N_A is known is fully treated in SSW. It is assumed that any element of A included in a sample can be identified. The problem is to construct a confidence interval for the domain total, $T_A = \sum_{i \in A} x_i$, based on a sample of n elements selected from the entire frame.

Explicitly (as in SSW, Section 3.3) or implicitly (as in SSW, Section 10.3) the standard approach to this problem is to redefine x_i , by setting $x_i = 0$ if $i \notin A$, which forces the population total $T = \sum_{i=1}^N x_i$ to be equal to T_A . Thus, the construction of a confidence interval for the domain total is recast as the construction of a confidence interval for a population total. In what follows it is assumed that the x_i 's have been redefined as above. We shall also assume, here and throughout this paper, that n is sufficiently large and n/N sufficiently small that second order terms can be ignored. Define the additional population parameters,

$\bar{X} = T/N$ = population mean,

$S^2 = \sum_{i=1}^N (x_i - \bar{X})^2/N$ = population variance, and

$p_A = N_A/N$ = proportion of population in A .

Then

- (1) $\hat{T}_A = (N/n) \sum_{i=1}^n x_i$, $\bar{x} = \sum_{i=1}^n x_i/n = \hat{T}_A/N$, $s^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$, and $\hat{p}_A = n_A/n$ (where n_A is the number of sample elements in A) are unbiased for the corresponding population parameters,
- (2) $E(\hat{T}_A) = T_A$,
- (3) $\text{var}(\hat{T}_A) = N^2 S^2/n$,
- (4) $\sqrt{n}(\hat{T}_A - T_A)/(NS) \xrightarrow{d} N(0, 1)$, and
- (5) s^2 is consistent for S^2 .

It follows that $\sqrt{n}(\hat{T}_A - T_A)/(Ns) \xrightarrow{d} N(0, 1)$, so, when n is "sufficiently large", appropriate values from the normal distribution can be used to construct confidence intervals for T_A , as noted by SSW, p. 391.

The proportion of the population in A^c is $1 - p_A$ and $x_i = 0$ for $i \in A^c$; therefore, when p_A is small and the values of the x_i 's for $i \in A$ are concentrated away from zero, the convergence in distribution in (4) can be slow.

Consequently, the distribution of $\sqrt{n}(\hat{T}_A - T_A)/Ns$ can deviate from normal even for what are usually considered to be moderate to large values of n . The simulation study in Section 2.5 illustrates this.

For the case of stratified random sampling, confidence interval coverage for domain quantities using standard methods can be poor. Dorfman and Valliant (1993) noted the problem in their study of wage distributions for domains consisting of workers in specific occupational groups. Preliminary empirical work by the authors indicated that supposed 95% confidence intervals for total workers and total wages for occupation based domains typically provided only 75% to 85% coverage even for a large total sample size ($n = 353$ establishments). These results are verified as part of the empirical work described in Section 3. Furthermore, their work indicated that the distribution of $\hat{T}_A - T_A$ was strongly dependent on the realized value of n_A , which suggested that some type of "conditional" confidence interval should be considered. It seems desirable to establish methodology for the construction of conditional (on n_A or equivalently \hat{p}_A) confidence intervals for T_A , which provide nominal, or near nominal, coverage regardless of the realized value of the domain sample size. Inference conditional on sample size is discussed in SSW, Section 10.4, but only for the case of known N_A ; we are concerned throughout this paper with the case of unknown N_A .

2.2 Definitions and Notation

We define the following parameters and estimators:

Domain parameters:

$\mu_A = T_A/N_A$ = domain mean,

$\sigma_A^2 = \sum_{i \in A} (x_i - \mu_A)^2/N_A$ = variance of population elements in A .

Domain estimators:

$\hat{N}_A = \hat{p}_A N$,

$\hat{\mu}_A = \sum_{i=1}^{n_A} x_i/n_A = \hat{T}_A/\hat{N}_A$ (only defined for $n_A \geq 1$), and

$\hat{\sigma}_A^2 = \sum_{i=1}^{n_A} (x_i - \hat{\mu}_A)^2/(n_A - 1)$ (only defined for $n_A \geq 2$).

In what follows it is understood that $n_A \geq 2$ (or equivalently $\hat{p}_A \geq 2/n$) unless specifically stated otherwise. At $n_A = 1$ or 0, it is preferable to supply an "insufficient information" tag, rather than attempt inference. The relationships given below follow directly from the definitions:

$T_A = N p_A \mu_A$ and $\hat{T}_A = N \hat{p}_A \hat{\mu}_A$,

$\bar{X} = p_A \mu_A$ and $\bar{x} = \hat{p}_A \hat{\mu}_A$,

$S^2 = p_A(1 - p_A)\mu_A^2 + p_A\sigma_A^2$

and

$$s^2 = \frac{n}{n-1} \hat{p}_A (1 - \hat{p}_A) \hat{\mu}_A^2 + \frac{n \hat{p}_A - 1}{n-1} \hat{\sigma}_A^2. \quad (1)$$

Also, it is straightforward to verify that

$$(\sqrt{n}/N)(\hat{T}_A - T_A) = \sqrt{n}\mu_A(\hat{p}_A - p_A) + \sqrt{\hat{p}_A}\sigma_A Z, \quad (2)$$

where $Z = \sqrt{n\hat{p}_A}(\hat{\mu}_A - \mu_A)/\sigma_A$. Thus, conditionally on \hat{p}_A , \hat{T}_A is biased for T_A ; and if, for example, we assume an underlying normality, and standardize $(\sqrt{n}/N)(\hat{T}_A - T_A)$ by the corresponding conditional variance, we will get a non-central t -distribution with unknown non-centrality parameter proportional to $\sqrt{n}\mu_A(\hat{p}_A - p_A)$, providing little basis for (conditional) sound inference. This is the problem which the discussions in the next sections attempt to address.

We remark that in estimating the mean μ_A by $\hat{\mu}_A$, the bias is zero, and the problem of the preceding paragraph does not arise. This is the reason that, in simple random sampling, standard inference for means is sound, at least when the domain variates are normally distributed.

2.3 General Methodology for Confidence Intervals

Let $\hat{\theta} = (\hat{T}_A - T_A)/s_{\hat{T}_A}$, where $s_{\hat{T}_A}^2$ is an estimator (to be specified) of the (conditional or unconditional) variance of the total. Assume that the form of the conditional (on \hat{p}_A) distribution function of $\hat{\theta}$, say $H(\cdot | \hat{p}_A; p_A, \mu_A, \sigma_A^2)$, is known where p_A, μ_A and σ_A^2 represent unknown parameters. In order to construct a conditional equal tailed $(1 - \alpha) \times 100\%$ confidence interval (CI) for T_A , we define an upper critical value

$$c_u \equiv c_u(\alpha, \hat{p}_A, p_A) = -\inf\{x | H(x | \hat{p}_A; p_A) \geq \alpha/2\} = -H^{-1}(\alpha/2, \hat{p}_A; p_A)$$

where p_A is considered fixed and the dependence on μ_A and σ_A^2 is temporarily suppressed; a lower critical value, say c_l , is defined in a similar manner. A conditional, equal tailed $(1 - \alpha) \times 100\%$ CI for T_A is then given by $CI(1 - \alpha) = (\ell, u)$, where

$$u = \hat{T}_A + c_u s_{\hat{T}_A} \text{ and } \ell = \hat{T}_A + c_l s_{\hat{T}_A}. \quad (3)$$

At this point the obvious practical problem is that the critical values c_u and c_l depend not only on \hat{p}_A but also on the unknown parameter p_A . One approach to this problem is to take a Bayesian tack and assume the parameter p_A is the realization of a random variable. Adjusting the notation to reflect the assumption that p_A is stochastic, we replace $H(x | \hat{p}_A; p_A)$ by $H(x | \hat{p}_A, p_A)$ and have that

$$\begin{aligned} \Pr\{\hat{\theta} \leq x | \hat{p}_A\} &= F(X | \hat{p}_A) \\ &= \frac{1}{h(\hat{p}_A)} \int H(x | \hat{p}_A, p_A) f(\hat{p}_A | p_A) g(p_A) dp_A, \end{aligned} \quad (4)$$

where $h(\hat{p}_A) = \int f(\hat{p}_A | p_A) g(p_A) dp_A$ and $g(p_A)$ is the density of p_A . It should be noted that as a consequence of our sampling scheme the distribution of $n\hat{p}_A$, conditional on p_A , is Binomial (n, p_A) so that $f(\hat{p}_A | p_A)$ is known. Under the Bayesian approach, the critical values are $c_u^* \equiv c_u^*(\alpha, \hat{p}_A) = -F^{-1}(\alpha/2 | \hat{p}_A)$ and $c_l^* \equiv c_l^*(\alpha, \hat{p}_A) = -F^{-1}(1 - \alpha/2 | \hat{p}_A)$ so the upper and lower limits for a conditional $(1 - \alpha) \times 100\%$ CI for T_A are

$$u = \hat{T}_A + c_u^* s_{\hat{T}_A} \text{ and } \ell = \hat{T}_A + c_l^* s_{\hat{T}_A}. \quad (5)$$

For the purposes of our current research, we assume that the prior distribution $g(p_A)$ is $N(\mu_{p_A}, \sigma_{p_A}^2)$ with μ_{p_A} and $\sigma_{p_A}^2$ to be specified, with the understanding that $\sigma_{p_A}^2$ is sufficiently small that p_A lies between 0 and 1 with near certainty. The normality assumption is made for mathematical convenience. It also captures notions we may have of degrees of closeness to, and symmetry about, μ_{p_A} . For an empirical Bayes approach, we use $\mu_{p_A} = \hat{p}_A$; we consider several possible alternatives for $\sigma_{p_A}^2$ discussed in detail below. Our experience indicates that the normality assumption is not crucial; rather, it is primarily a matter of convenience.

2.4 Confidence Intervals Under Normal Assumptions

To proceed further we assume that within the domain A the x_i are distributed $N(\mu_A, \sigma_A^2)$. In practice, this assumption may not be met. Nonetheless, it leads to suggested modifications that will not at any rate give lower coverage of confidence intervals than the standard approach. Combining this assumption with earlier results, in particular equation (2), and ignoring lower order terms, we have

- (a) $[\sqrt{n}(\hat{T}_A - T_A)/n | \hat{p}_A, p_A]$ is distributed $N(\sqrt{n}\mu_A(\hat{p}_A - p_A), \hat{p}_A\sigma_A^2)$,
- (b) $\left[(n\hat{p}_A - 1) \frac{\hat{\sigma}_A^2}{\sigma_A^2} | \hat{p}_A, p_A \right]$ is distributed $\chi^2(n\hat{p}_A - 1)$, and
- (c) the conditional random variable in (b) is stochastically independent of the conditional random variable in (a).

Consider $\hat{\theta}_1 = (\hat{T}_A - T_A)/(N\hat{\sigma}_A\sqrt{\hat{p}_A}/\sqrt{n})$, which utilizes the conditional variance of \hat{T}_A as the standardizing term. It follows immediately from (a), (b) and (c) that, conditional on (\hat{p}_A, p_A) the random variable $\hat{\theta}_1$ is distributed as a non-central t with $n\hat{p}_A - 1 = n_A - 1$ degrees of freedom and non-centrality parameter

$$\lambda = \sqrt{n}\gamma_A(\hat{p}_A - p_A)/\sqrt{\hat{p}_A},$$

with

$$\gamma_A = \mu_A/\sigma_A.$$

Thus, we have specified the conditional distribution function $H(\cdot | \hat{p}_A, p_A)$ of $\hat{\theta}_1$. As $f(\hat{p}_A | p_A)$ and $g(p_A)$ have been previously specified, it follows that $F(\cdot | \hat{p}_A)$ in (4) is well-defined although extremely cumbersome to calculate. The dependence on μ_A and σ_A^2 , through γ_A , should be noted.

Although $F(\cdot | \hat{p}_A)$ as given above can be used to determine the critical values, they are extremely difficult to calculate. A relatively simple approach, given in the next paragraph, provides a close approximation to the critical values. We have verified the closeness of the approximation by computing the exact values for selected cases using large scale simulations.

Adoption of a locally uniform prior on p_A leads to the approximate posterior distribution $p_A \sim N(\hat{p}_A, \text{var}(\hat{p}_A))$ and we could approximate $\text{var}(\hat{p}_A)$ by $\hat{p}_A(1 - \hat{p}_A)/n$. We adopt the slightly more flexible prior $p_A \sim N(\mu, \sigma_{p_A}^2)$, and empirically choose $\mu = \hat{p}_A$, with several possibilities for $\sigma_{p_A}^2$ that will be specified below. It follows from Appendix A that $[\lambda | \hat{p}_A]$ is distributed approximately as a normal with mean zero and variance $\gamma_A^2(1 - \hat{p}_A)/(1 + \psi_A)$, where

$$\psi_A = \hat{p}_A(1 - \hat{p}_A)/n\sigma_{p_A}^2.$$

Then, from the result in Appendix B, conditional on \hat{p}_A ,

$$\frac{(\hat{T}_A - T_A)}{\frac{N\hat{\sigma}_A\sqrt{\hat{p}_A}}{\sqrt{n}} \sqrt{\frac{\gamma_A^2(1 - \hat{p}_A)}{1 + \psi_A} + 1}}$$

is distributed as a central t with $n_A - 1$ degrees of freedom. Let $t_{1-\alpha/2, n_A-1}$ be the $(1 - \alpha/2)100\%$ percentile of this distribution. The upper confidence limit u , defined in (5), is given (approximately) by

$$u = \hat{T}_A + N\hat{\sigma}_A\sqrt{\hat{p}_A/n} \times \left(\left(\gamma_A^2(1 - \hat{p}_A) + 1 + \psi_A \right) / (1 + \psi_A) \right)^{1/2} t_{1-\alpha/2, n_A-1}. \quad (6)$$

As $\hat{\sigma}_A^2$ is conditionally unbiased for σ_A^2 and $\hat{\mu}_A^2 - \hat{\sigma}_A^2/n_A$ is conditionally unbiased for μ_A^2 , we use $\hat{\gamma}_A^2 = (\hat{\mu}_A^2 - \hat{\sigma}_A^2/n_A)/\hat{\sigma}_A^2$ to estimate γ_A^2 . Substituting $\hat{\gamma}_A^2$ for γ_A^2 in (6) yields

$$\tilde{u} \approx \hat{T}_A + (Ns/\sqrt{n}) \times \left(\left(1 + \frac{\hat{p}_A\hat{\sigma}_A^2\psi_A}{s^2} \right) / (1 + \psi_A) \right)^{1/2} t_{1-\alpha/2, n_A-1} \quad (7)$$

where s^2 is defined in (1).

It remains to choose ψ_A . We note that \tilde{u} is strictly decreasing as ψ_A increases and

$$\tilde{u} \rightarrow \hat{T}_A + \frac{Ns}{\sqrt{n}} t_{1-\alpha/2, n_A-1} = \tilde{u} \text{ as } \psi_A \text{ becomes small,}$$

$$\tilde{u} = \hat{T}_A + \frac{Ns}{\sqrt{n}} \left(\frac{1 + \hat{p}_A\hat{\sigma}_A^2/s^2}{2} \right)^{1/2} t_{1-\alpha/2, n_A-1} = \tilde{u}_2 \text{ for } \psi_A = 1,$$

and

$$\tilde{u} \rightarrow \hat{T}_A + \frac{Ns}{\sqrt{n}} \left(\frac{\sqrt{\hat{p}_A}\hat{\sigma}_A}{s} \right) t_{1-\alpha/2, n_A-1} = \tilde{u}_3 \text{ as } \psi_A \text{ becomes large.} \quad (8)$$

In each case the lower critical value can be dealt with in an analogous manner resulting in three competing confidence intervals; namely, $CI_i(1 - \alpha) = (\tilde{\ell}_i, \tilde{u}_i)$, $i = 1, 2, 3$, with $\tilde{\ell}_i$ defined similarly to \tilde{u}_i in (8) with $t_{1-\alpha/2, n_A-1}$ replaced by $t_{\alpha/2, n_A-1}$. The competing confidence intervals are labeled in order of decreasing length.

The first case is equivalent to assuming that $\sigma_{p_A}^2$ is large relative to $\text{var}(\hat{p}_A)$ and leads to using the usual unconditional variance but with degrees of freedom equal to $n_A - 1$. In most practical problems this seems reasonable since $\sigma_{p_A}^2$ is an unknown constant and $\text{var}(\hat{p}_A)$ is $O(p_A/n)$. The second interval corresponds to adoption of a normal prior as noted above, with $\sigma_{p_A}^2 = \hat{p}_A(1 - \hat{p}_A)/n$. The last confidence interval is based on the assumption that p_A is essentially degenerate at \hat{p}_A .

2.5 Empirical Study for SRS

We compared the several confidence intervals of Section 2.4 in a small empirical study, using artificial populations, for which the domain variable was normal. In all cases the population size N was 1,000, and the sample size n was 100 or 300. The parameters p_A and γ_A varied from population to population. Letting M_2 be the number of runs with $n_A \geq 2$, we allowed the run size M to vary to give $M_2 = 10,000$. Table 1 gives coverage results. CI_0 represents the confidence interval based on the standard normal methodology. The results for CI_2 closely approximated the results for CI_1 and are excluded. The value of M is included to indicate how many trials fell into the "insufficient information" pile, at a given setting of the parameters. Several conclusions seem warranted:

1. Standard confidence intervals using the usual variance estimate and normal quantiles can give low coverage. This occurs for several values of p_A when $\gamma_A = 1/2$ or $\gamma_A = 2$, however, the under-coverage is not too severe if the domain variable is normal. The case where

- $\gamma_A = 2$ or takes even larger values is probably more likely in practice. Thus if the domain variable is normal, the use of standard confidence intervals under simple random sampling case is not particularly worrisome.
2. The strictly conditional intervals (*i.e.*, CI_3) using the conditional variance can give abominable coverage, when γ_A is large. That is, confidence intervals based on "large" values of ψ_A gave very poor results.
 3. The use of the standard variance estimate but replacing the standard normal quantile with a t -quantile having degrees of freedom based on the number of sample units in the domain (*i.e.*, CI_1) gives approximately nominal or conservative coverage regardless of the value of γ_A .

Table 1

Coverage of 95% Confidence Intervals for Domain Total for Artificial Populations with Domain Variate Normally Distributed*

			Coverage		
P_A	n	M	CI_0	CI_1	CI_3
$y = 1/2$					
.01	100	38774	100.0	100.0	91.2
	300	11773	98.3	100.0	83.2
.02	100	16327	91.1	99.4	95.0
	300	10078	88.6	95.5	93.9
.05	100	10303	88.7	97.8	93.5
	300	10000	92.3	94.4	92.5
.10	100	10001	90.9	94.8	92.5
	300	10000	94.0	95.0	92.3
$y = 2$					
.01	100	37749	99.9	100.0	83.5
	300	11740	94.4	100.0	89.1
.02	100	16348	99.0	100.0	88.4
	300	10075	91.4	98.9	74.7
.05	100	10312	90.5	99.5	77.6
	300	10000	93.8	95.8	66.6
.10	100	10000	91.7	96.5	67.9
	300	10000	94.0	95.2	65.0

* See Equation (8) and accompanying text for definition of CI_1 and CI_3 . CI_0 is the standard normal confidence interval.

As a minor observation on the results, we note the counter-intuitive increases in coverage for smaller p_A and n . We believe this is due to the fact that, at very small values of p_A and n , \hat{p}_A is constrained to be positive, and so cannot deviate much below p_A . Were intervals calculable for $n_A = 0$, there would be a serious drop in coverage in these cases. Note that the coverage rises unexpectedly only where M is large.

3. THE CASE OF STRATIFIED RANDOM SAMPLING

3.1 Definitions and Notation

Assume there are K strata and, where appropriate, terms previously defined have corresponding stratum level

definitions. For example, n_k is the sample size and n_{Ak} is the number of sample elements in A for the k -th stratum. Thus, a natural estimator for the domain total

$$T_A = \sum_{k=1}^K \sum_{i \in A} x_{ki} = \sum_{k=1}^K N_k \hat{p}_{Ak} \mu_{Ak} \text{ is}$$

$$\hat{T}_A = \sum_{k \in B_1} \hat{T}_{Ak} = \sum_{k \in B_1} N_k \hat{p}_{Ak} \hat{\mu}_{Ak},$$

where $\hat{p}_{Ak} = n_{Ak}/n_k$, $\hat{\mu}_{Ak} = \sum_{i=1}^{n_{Ak}} x_{ki}/n_{Ak}$ and $B_1 = \{k | n_{Ak} \geq 1 \text{ and } 1 \leq k \leq K\}$. As $\hat{p}_{Ak} = 0$ for $k \notin B_1$, it is straightforward to verify that

$$E[(\hat{T}_A - T_A) | \hat{p}_A, p_A] = \sum_{k=1}^K N_k (\hat{p}_{Ak} - p_{Ak}) \mu_{Ak} \equiv \tilde{\mu}_A \quad (9)$$

and

$$\begin{aligned} \text{var}[(\hat{T}_A - T_A) | \hat{p}_A, p_A] &= \sum_{k \in B_1} N_k^2 \hat{p}_{Ak}^2 \sigma_{Ak}^2 / n_{Ak} = \\ &\sum_{k \in B_1} N_k^2 \hat{p}_{Ak}^2 \sigma_{Ak}^2 / n_k \equiv \tilde{\sigma}_A^2, \end{aligned}$$

where $\hat{p}_A = [\hat{p}_{A1} \hat{p}_{A2} \dots \hat{p}_{AK}]$, $p_A = [p_{A1} p_{A2} \dots p_{AK}]$. Thus, as in the simple random sampling case, there is a conditional bias $\tilde{\mu}_A$, which needs to be taken into account.

3.2 A Methodology for Confidence Intervals

The general methodology for confidence intervals of Section 2.3 for simple random sampling holds here as well. One need only reinterpret scalars as vectors; for example, replace \hat{p}_A by $\hat{p}_A = (\hat{p}_{A1}, \dots, \hat{p}_{AK})'$. In particular, $H(x | \hat{p}_A, p_A) = \Pr\{\hat{\theta} \leq x | \hat{p}_A, p_A\}$ will be the conditional distribution function of $\hat{\theta} = (\hat{T}_A - T_A) / \hat{\sigma}_A$, where $\hat{\sigma}_A$ is a re-scaling factor to be specified.

Let $B_2 = \{k | n_{Ak} \geq 2 \text{ and } 1 \leq k \leq K\}$ and, for $k \in B_2$, define $\hat{\sigma}_{Ak}^2 = \sum_{i=1}^{n_{Ak}} (x_{ki} - \hat{\mu}_{Ak})^2 / (n_{Ak} - 1)$. Under normality, $(n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2 \sim \chi^2(n_{Ak} - 1)$, so if $\{d_k | k \in B_2\}$ are non-negative constants with $\sum_{k \in B_2} d_k > 0$, then by the usual Satterthwaite (1946) two moment approximation, the conditional random variable

$$\left[(1/c) \sum_{k \in B_2} d_k (n_{Ak} - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2) | \hat{p}_A, p_A \right]$$

is distributed approximately as a $\chi^2(v)$, where

$$c = \sum_{k \in B_2} d_k^2 (n_{Ak} - 1) / \sum_{k \in B_2} d_k (n_{Ak} - 1)$$

and

$$v = (\sum_{k \in B_2} d_k (n_{Ak} - 1))^2 / \sum_{k \in B_2} d_k^2 (n_{Ak} - 1).$$

This suggests that we restrict our attention to expressions of the general form

$$\hat{\sigma}_A^2 = \sum_{k \in B_2} d_k (n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$$

with choice of the d_k to be specified. Note that when $B_1 = B_2$ and $d_k = N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 / n_k (n_{Ak} - 1)$, $\hat{\sigma}_A^2 = \hat{\sigma}_A^2 \equiv \sum_{k \in B_2} d_k (n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$ is an unbiased estimator for the conditional variance $\hat{\sigma}_A^2$. However, as in the simple random sampling case, this estimator will tend to be too small. We use the more general expression to develop a family of t -statistics when we “uncondition” on \mathbf{p}_A . Each of these will involve unknown parameters, and, as in the simple random sampling case (transition of equation (6) to equation (7)), estimation of these unknowns will be necessary. Thus the net result will be several rival “near t -statistics” which we may then compare empirically.

Because the samples are selected independently from each stratum we have $f(\hat{\mathbf{p}}_A | \mathbf{p}_A) = \prod_{k=1}^K f_k(\hat{p}_{Ak} | p_{Ak})$ and, as a consequence of our within stratum sampling scheme, $n_k \hat{p}_{Ak}$ has a binomial distribution $B(n_k, p_{Ak})$. We assume that the $\{p_{Ak} | 1 \leq k \leq K\}$ are jointly independent so $g(\mathbf{p}_A) = \prod_{k=1}^K g_k(p_{Ak})$ which implies

$$f(\hat{\mathbf{p}}_A | \mathbf{p}_A) g(\mathbf{p}_A) = \prod_{k=1}^K f_k(\hat{p}_{Ak} | p_{Ak}) g_k(p_{Ak})$$

and

$$h(\hat{\mathbf{p}}_A) = \prod_{k=1}^K \int f_k(\hat{p}_{Ak} | p_{Ak}) g_k(p_{Ak}) dp_{Ak}.$$

In what follows, we assume that the prior distribution of p_{Ak} is $N(\mu_{p_{Ak}}, \sigma_{p_{Ak}}^2)$ and for the empirical Bayes approach, we use $\mu_{p_{Ak}} = \hat{p}_{Ak}$ and, analogously to the case of simple random sampling, we define

$$\psi_{Ak} = \hat{p}_{Ak}(1 - \hat{p}_{Ak}) / n_k \sigma_{p_{Ak}}^2.$$

It is straightforward to extend the result in Appendix A to the case of stratified random sampling and it then follows that, for $\tilde{\mu}_A$ defined by (9), $[\tilde{\mu}_A / \tilde{\sigma}_A | \hat{\mathbf{p}}_A]$ is distributed $N(0, \text{var}(\tilde{\mu}_A | \hat{\mathbf{p}}_A) / \tilde{\sigma}_A^2)$, where $\text{var}(\tilde{\mu}_A | \hat{\mathbf{p}}_A) = \sum_{k \in B_1} N_k^2 \mu_{Ak}^2 \hat{p}_{Ak}(1 - \hat{p}_{Ak}) / n_k(1 + \psi_{Ak})$. Using the result in Appendix B, it follows that, conditional on $\hat{\mathbf{p}}_A$, the random variable

$$\hat{\Theta} = \frac{(\hat{T}_A - T_A) / \sqrt{\text{var}(\tilde{\mu}_A | \hat{\mathbf{p}}_A) + \tilde{\sigma}_A^2}}{\sqrt{\hat{\sigma}_A^2 / cv}} = \frac{(\hat{T}_A - T_A) / \sqrt{\text{var}(\tilde{\mu}_A | \hat{\mathbf{p}}_A) + \tilde{\sigma}_A^2}}{\sqrt{\sum_{k \in B_1} d_k (n_{Ak} - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2) / \sum_{k \in B_1} d_k (n_{Ak} - 1)}}$$

is distributed approximately as a central t with v degrees of freedom.

Letting $\Theta = \text{var}(\tilde{\mu}_A | \hat{\mathbf{p}}_A) + \tilde{\sigma}_A^2$, with

$$\gamma_{Ak}^2 = \mu_{Ak}^2 / \sigma_{Ak}^2$$

and assuming the ψ_{Ak} are near zero we have

$$\Theta = \sum_{k \in B_1} \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1).$$

Thus, the upper bound on the CI would be (approximately)

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} d_k (n_{Ak} - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2)}}{\sqrt{\sum_{k \in B_2} d_k (n_{Ak} - 1)}} \Theta^{1/2} t_v, \quad (10)$$

where t_v stands for the critical values of the t_v distribution. Unfortunately the bound depends not only on our choice of the d_k , but also on the unknown parameters μ_{Ak} and σ_{Ak}^2 .

It is not hard to show that $v \leq \sum_{k \in B_2} (n_{Ak} - 1) \equiv v_{\max}$ and, if we set $d_k = 1$ (or any constant for that matter) then $v = v_{\max}$. We refer to v_{\max} specifically as the unweighted degrees of freedom. In this case the upper bound on the CI would be

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} d_k (n_{Ak} - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2)}}{\sqrt{\sum_{k \in B_2} (n_{Ak} - 1)}} \Theta^{1/2} t_{v_{\max}}.$$

Another approach is to attempt to finesse the problem of estimating Θ (at least when $B_1 = B_2$) by a judicious choice of the d_k . To that end let us assume that $B_1 = B_2$ and let

$$d_k = \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k (n_{Ak} - 1)} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)$$

so that $\sum_{k \in B_2} d_k (n_{Ak} - 1) = \Theta$ and Θ cancels out in (10). We then have

$$u = \hat{T}_A + \sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)} t_{v_1},$$

where v_1 is the degrees of freedom associated with this second choice of the d_k . More generally (*i.e.*, when $B_1 \neq B_2$), we have

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)}}{\sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2}{n_k} (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1)}} \Theta^{1/2} t_{v_1}.$$

In any event, we are still faced with the problem of estimating the population parameters and we have the additional problem of estimating the degrees of freedom.

A third possibility, which we have already mentioned, is to let $d_k = N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 / n_k (n_{Ak} - 1)$ so that when $B_1 = B_2$, $\hat{\sigma}_A^2 = \hat{\sigma}_A^2 = \sum_{k \in B_2} d_k (n_{Ak} - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$ is a conditionally unbiased estimator for σ_A^2 . In this case we have

$$u = \hat{T}_A + \frac{\sqrt{\sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 / n_k}}{\sqrt{\sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 / n_k}} \Theta^{1/2} t_{v_2},$$

where v_2 is the degrees of freedom associated with this third choice of the d_k . As in the second case, we are faced with the problem of estimating the population parameters and the degrees of freedom.

Now, it should be noted that if we estimate σ_{Ak}^2 with $\hat{\sigma}_{Ak}^2$ for $k \in B_2$ and let $\hat{\Theta}$ be a yet to be specified estimator of Θ then the (estimated) upper bounds above are $u = \hat{T}_A + \hat{\Theta}^{1/2} t_{v_{\max}}$, $u = \hat{T}_A + \hat{\Theta}^{1/2} t_{\hat{v}_1}$ and $u = \hat{T}_A + \hat{\Theta}^{1/2} t_{\hat{v}_2}$ respectively. The degrees of freedom are estimated by substituting estimates of the population parameters into the two respective choices of the d_k . Both \hat{v}_1 and \hat{v}_2 are smaller than v_{\max} , so, for any realized value of $\hat{\Theta}$, the confidence interval using v_{\max} will be the shortest. There is no general relationship between the sizes of \hat{v}_1 and \hat{v}_2 . Empirical evidence indicates that there is little to choose between the second and third approach.

Addressing the problem of estimating Θ , we can write

$$\Theta = \sum_{k \in B_1 - B_2} N_k^2 \hat{p}_{Ak} (\mu_{Ak}^2 (1 - \hat{p}_{Ak}) + \sigma_{Ak}^2) / n_k + \sum_{k \in B_2} N_k^2 \hat{p}_{Ak} (\mu_{Ak}^2 (1 - \hat{p}_{Ak}) + \sigma_{Ak}^2) / n_k.$$

For $k \in B_1 - B_2$ the estimator $\hat{\sigma}_{Ak}^2$ is not defined, however, it is straightforward to verify that $(1 - \hat{p}_{Ak}) E[\hat{\mu}_{Ak}^2 | n_{Ak}] \leq \sigma_{Ak}^2 + \mu_{Ak}^2 (1 - \hat{p}_{Ak}) \leq E[\hat{\mu}_{Ak}^2 | n_{Ak}]$. It follows that

$$s_a^2 = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) \hat{\mu}_{Ak}^2 / n_k + \sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 (1 + 1/n_k - 1/n_{Ak}) / n_k$$

will tend to underestimate Θ , and

$$s_b^2 = \sum_{k \in B_1 - B_2} N_k^2 \hat{p}_{Ak} \hat{\mu}_{Ak}^2 / n_k + \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) \hat{\mu}_{Ak}^2 / n_k + \sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 (1 + 1/n_k - 1/n_{Ak}) / n_k$$

will tend to overestimate Θ . Clearly, $s_a^2 \leq s_b^2$ with equality only when $B_1 = B_2$.

It can also be verified that in the case of stratified sampling, the standard variance estimator for estimated population totals is

$$s_{\text{std}}^2 = \sum_{k \in B_1} N_k^2 s_k^2 / n_k = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} (1 - \hat{p}_{Ak}) \hat{\mu}_{Ak}^2 / (n_k - 1) + \sum_{k \in B_2} N_k^2 \hat{p}_{Ak} \hat{\sigma}_{Ak}^2 (1 - 1/n_{Ak}) / (n_k - 1).$$

This looks like a satisfactory estimator of Θ , if the n_k are not small.

These results imply that CIs of the form $(\hat{T}_A \pm s_b t_{1-\alpha/2, v_1})$ will provide the highest level of coverage; but CIs of the form $(\hat{T}_A \pm s_{\text{std}} t_{1-\alpha/2, v_{\max}})$ and even perhaps $(\hat{T}_A \pm s_{\text{std}} t_{1-\alpha/2, v_1})$ have obvious computational advantages. Several of these competing forms of CI are evaluated empirically in Section 3.3. These results can easily be extended to ratio estimators by the standard linearization approach.

3.3 Empirical Investigation for Stratified Random Sampling: the BLS Wage Data

With a view to improving estimation of precision on wage data produced by the U.S. Bureau of Labor Statistics, we investigated coverage and interval length in two simulation studies on populations constructed from a test sample of the Occupational Compensation Survey Program (OCSF) conducted in 1991. The OCSF consisted of establishment surveys in several metropolitan areas, aimed at estimating wages levels for a select group of occupations. The surveys were carried out by stratified simple random sampling, with establishments stratified by employment size and industrial classification.

One population (the "Small Population") took the test sample itself as the population, with six non-certainty strata, and one certainty stratum of 12 establishments. Five hundred stratified random samples were taken from this population at sizes $n = 36$ and 60 , corresponding to the choices $n_k = 4$ and $n_k = 8$, reflecting relative sample sizes of sampling from the original population. The second population (the "Large Population") was constructed by expanding the sample data through replication (by simple random sampling with replacement, within each Small Population stratum) of establishments to achieve a population the size of the original population; again there were six noncertainty and one certainty strata; for each stratum sample sizes were the same as in the actual sample. Domains are defined by the different occupations of interest; only a fraction of establishments have workers in a particular occupation, and lie in the corresponding domain. Table 2 gives the number of establishments having workers in the selected occupations for the small population.

In both cases sampling was without replacement, so finite population correction factors were included (as appropriate) in the construction of the CIs. Also, the study was limited to a concern with 95% coverage.

Table 2
Number of Establishments in Given Domain (Occupation),
by Stratum for Small Population

Occupation	stratum							total
	1	2	3	4	5	6	7	
4021	0	4	11	10	8	10	7	50
1141	0	3	11	7	11	9	7	48
1122	0	3	8	13	14	12	6	56
3180	10	11	5	25	20	4	5	80
2911	0	3	14	2	13	17	7	56
1142	2	8	15	9	15	19	9	77
1180	17	20	5	61	31	3	1	138
1403	12	16	22	28	25	27	9	139
All Estabs	35	35	33	136	66	36	12	353

Small Population: Table 3 gives coverage and median relative interval length for total wages, at two sample sizes $n_k = 4$ and $n_k = 8$, for 8 occupations, and three methods of confidence interval construction: the standard variance estimator, s_{std}^2 , with the standard normal z -quantile, the unweighted degrees of freedom v_{max} , and the weighted degrees of freedom v_1 . Occupations are ordered by increasing values of the average value, over runs, of the unweighted degrees of freedom. We note:

- 1) Almost universally, coverage using the standard variance estimator and the standard normal quantiles (infinite df) is poor.
- 2) Coverage for the other interval types is far more satisfactory. In general, the coverage is near the nominal 95%, or slightly conservative, for weighted degrees of freedom; as expected, intervals based on unweighted degrees of freedom tend to yield coverage a few points below those based on weighted degrees of freedom.

- 3) Two occupations (1122, 4021) yield seriously low coverage for totals even with the improved procedures. Investigation of these particular occupations suggests a strong violation of the normality assumption. In 4021, for example, two units in stratum 5 have a number of workers, and hence total wages, an order of magnitude higher than the other establishments in this stratum and indeed in the population. Furthermore, the wage rate of these two outliers is markedly lower than the great bulk of establishments: with just these two excluded from the population, the overall population average wage would be \$9.68/hour; with them in, it is \$8.28. Since there are 66 establishments in stratum 5, it is easy for these two establishments to escape being in a sample of size 8; the consequence is a serious overestimate of the mean wage or underestimate of total wage. At the same time, wages for the establishments that are in the sample are relatively homogeneous, so the variance estimate will tend to be too low. The presence of several smaller establishments in the domain contribute to enlarging the degrees of freedom, and so the t -adjustment is unable to compensate fully. It is hard to see how to guard against such a problem short of having prior information, and allotting such outliers to a certainty stratum. Even so, the adjusted intervals are a significant improvement on the naïve normal distribution based interval.

Interval lengths are taken relative to $2 \times z_{.975} \approx 4$ times the root mean square error of \hat{T}_A calculated over runs. We report the median of these standardized lengths (across runs). When the distribution of \hat{T}_A is actually normal, the median length is close to 1.

Table 3
Estimated degrees of freedom, coverage, and relative median length of CIs for total wages of workers in occupation,
for the small population

Occupation	Four Sample Establishments Per Stratum								Eight Sample Establishments Per Stratum							
	4021	1141	1122	3180	2911	1142	1180	1403	1141	4021	1122	3180	2911	1142	1180	1403
$df = v_{\text{max}}$	1.5	1.6	1.6	2.0	2.3	2.8	4.3	6.1	3.7	3.8	3.9	5.6	6.0	8.0	12.3	16.6
$df = \hat{v}_1$	1.3	1.3	1.4	1.5	1.7	1.9	2.3	3.5	2.0	2.3	2.3	3.1	3.5	4.3	5.4	9.7
Coverage																
$\hat{T}_A \pm s_{\text{std}} z$.47	.69	.51	.75	.73	.85	.89	.87	.74	.49	.65	.79	.78	.86	.88	.92
$\hat{T}_A \pm s_{\text{std}} t_{v_{\text{max}}}$.89	.92	.93	.99	.95	.96	.97	.92	.87	.65	.75	.89	.86	.90	.90	.94
$\hat{T}_A \pm s_{\text{std}} t_{\hat{v}_1}$.92	.93	.95	.99	.96	.96	.98	.95	.91	.74	.80	.94	.89	.95	.96	.96
Median Relative Length																
$\hat{T}_A \pm s_{\text{std}} z$	0.53	0.75	0.59	0.70	0.74	0.85	0.90	0.88	0.87	0.63	0.66	0.80	0.83	0.88	0.92	0.96
$\hat{T}_A \pm s_{\text{std}} t_{v_{\text{max}}}$	2.65	3.67	2.80	2.60	2.20	1.98	1.50	1.14	1.63	1.09	1.13	1.10	1.10	1.06	1.02	1.04
$\hat{T}_A \pm s_{\text{std}} t_{\hat{v}_1}$	3.30	4.32	3.19	3.40	3.08	3.06	2.70	1.58	3.08	2.40	2.38	2.00	1.74	1.38	1.38	1.13

- 4) The relative interval length of the standard interval tends to be too small, that is, it tends to be less than 1.
- 5) Interval length among the other variance-degrees of freedom combinations is largest for s_{std}^2 with \hat{v}_1 , and smallest for s_{std}^2 with v_{max} . These differences can be appreciable; there is a tradeoff between coverage and interval size.
- 6) For a given interval type, the relative interval length tends to 1 as v_{max} increases. The conclusions from a study of mean wages are similar.

Large Population: Table 4 gives coverage and interval length for total wages for five interval types, and a wider range of occupations, ordered by average v_{max} . The interval types include the three used previously for the small population. The two new intervals utilize the weighted degrees of freedom together with s_a and s_b respectively. Results are based on 5,000 runs.

- 1) The results are consistent with those for the Small Population, in terms of the relative coverage and interval sizes of the several interval types. The standard normal is unsatisfactory for many occupations.
- 2) The coverage for intervals using the weighted degrees of freedom, \hat{v}_1 , is less than 90% for only a small fraction of cases.

- 3) There can be marked differences in interval length for the different interval types; however, all ratios of interval length to $4 \times$ root mean square error tend to 1, as v_{max} gets large.
- 4) Little difference results from using s_a , s_b , or s_{std} with $t_{\hat{v}_1}$. Again, the results for mean wages, while differing in detail, lead to the same overall conclusions, and are omitted.

4. SUMMARY AND CONCLUSIONS

From our theoretical investigation and simulation work, we draw the following conclusions:

1. Standard 95% confidence intervals for domain means or totals, when based on the standard normal distribution and standard methods of variance estimation, tend to yield less than actual 95% coverage. The extent of the deviation will vary with domain (occupation in the wage study), but can be quite considerable even when the sample size is large.
2. New nonstandard methods offer a sharp improvement, giving intervals with better coverage, typically at or close to the nominal 95% coverage. These intervals tend to be longer than the standard intervals. The increase in length will vary with domain, and will depend on the particular method for CI construction that is adopted.

Table 4
Estimated degrees of freedom, coverage, and relative median length of CIs for total wages of workers in occupation, for the large population

	Occupation																				
	1718	1604	1802	1716	2911	2052	1332	1141	4021	1232	2853	3020	1122	1142	1714	1514	3180	4030	1063	1403	1180
$df = v_{\max}$	2.97	3.45	4.44	11.9	12.4	13.1	15.3	16.9	16.8	17.3	20.6	24.9	28.0	28.6	29.1	34.8	41.5	59.9	77.6	77.9	128
$df = \hat{v}_1$	2.67	2.34	2.35	5.97	5.90	4.25	11.4	9.00	6.32	15.5	13.5	10.4	15.2	9.67	15.3	18.0	25.2	14.3	27.4	28.5	90.0
	Coverage																				
$\hat{T}_A \pm s_{std} z$.89	.60	.85	.87	.87	.89	.93	.93	.89	.92	.92	.92	.88	.89	.85	.93	.92	.81	.94	.94	.94
$\hat{T}_A \pm s_{std} t_{v_{\max}}$.96	.83	.94	.89	.88	.91	.95	.95	.91	.94	.94	.93	.88	.90	.86	.93	.92	.81	.95	.94	.95
$\hat{T}_A \pm s_a t_{\hat{v}_1}$.97	.88	.94	.91	.89	.97	.96	.96	.91	.94	.94	.95	.89	.91	.86	.94	.93	.83	.95	.94	.95
$\hat{T}_A \pm s_{std} t_{\hat{v}_1}$.97	.89	.94	.92	.90	.97	.96	.91	.94	.94	.95	.89	.89	.91	.86	.94	.93	.83	.95	.95	.95
$\hat{T}_A \pm s_b t_{\hat{v}_1}$.97	.89	.97	.92	.90	.97	.96	.96	.91	.95	.94	.95	.89	.91	.87	.95	.93	.83	.95	.94	.95
	Median Relative Length																				
$\hat{T}_A \pm s_{std} z$	0.99	0.78	0.92	0.97	0.95	0.96	0.99	0.98	0.96	0.97	0.98	.98	0.95	0.96	0.93	0.98	1.00	0.91	1.00	1.00	1.01
$\hat{T}_A \pm s_{std} t_{v_{\max}}$	2.14	1.47	1.40	1.08	1.06	1.06	1.08	1.06	1.04	1.04	1.04	1.03	0.99	1.00	0.98	1.01	1.03	0.93	1.01	1.01	1.02
$\hat{T}_A \pm s_a t_{\hat{v}_1}$	2.32	2.24	2.46	1.37	1.37	1.59	1.12	1.15	1.34	1.05	1.11	1.16	1.04	1.19	1.04	1.04	1.05	1.07	1.09	1.04	1.02
$\hat{T}_A \pm s_{std} t_{\hat{v}_1}$	2.34	2.27	2.48	1.37	1.39	1.60	1.13	1.18	1.34	1.05	1.13	1.18	1.04	1.20	1.04	1.04	1.06	1.07	1.10	1.05	1.02
$\hat{T}_A \pm s_b t_{\hat{v}_1}$	2.47	2.33	2.79	1.39	1.38	1.61	1.14	1.20	1.35	1.07	1.13	1.18	1.04	1.19	1.05	1.05	1.06	1.07	1.10	1.04	1.02

For domains which yield large samples, there will be little difference from standard intervals.

3. The instances where coverage fell below nominal, even using the t -adjusted intervals, may be ascribed to severe violation of the normality assumption for the domain data. Thus the t -adjustment is not a cure-all. Nonetheless, even in such cases there is a good deal of improvement in coverage over the use of the standard normal interval.
4. The key idea behind these intervals is to condition on the amount of information on the particular occupation, which, roughly speaking, is measured in terms of the number of units in the sample that belong to the domain. The fraction of such units within each stratum is unknown, and to handle this fact we put a prior distribution on this unknown, reflective of the degree of our ignorance of it, an idea we borrow from the Bayesians. However, in the final analysis, it is the realized coverage probabilities that determine the merit of the approach.
5. The principal effect of these ideas is the abandonment, for purposes of CI construction, of the standard normal quantiles (± 1.96 for 95% coverage). These are replaced by quantiles from the Student's t -distribution, with degrees of freedom determined from the sample and varying with domain. If because of publication requirements or for other reasons, there is need to report standard deviations rather than confidence intervals, then we recommend reporting an effective standard deviation given by the length of the proposed t -based 95% confidence interval divided by twice 1.96.
6. The standard estimate of variance seems acceptable for estimating the variance, when accompanying the new t -quantile. In most instances this combination should be quite satisfactory, so that the only change from standard methodology will be the introduction of adjusted degrees of freedom. However, in some instances, the alternative standard deviations may improve coverage or reduce the length of confidence intervals.
7. An open question concerns what degree and type of collapsing of strata (if any) should be used in the estimation of variances and of the degrees of freedom for the purpose of confidence interval construction. In general, there will be a tradeoff: as strata are reduced in number, the estimate of variance will tend to increase, but so will the degrees of freedom (reducing the size of $t_{v_{\max}}$ or $t_{\hat{p}_i}$). The answer to this question may be population specific, and experience from past surveys useful.

ACKNOWLEDGEMENTS

The authors wish to thank the Associate Editor and the three referees for their many helpful suggestions. The research of S. Wang was supported in part by an ASA/NSF/BLS research fellowship, and grants from the National Security Agency (MDA904-96-1-0029), the National Science Foundation (DMS-9504589) and the Texas A&M University Scholarly and Creative Activities Program (95-59). The views are those of the authors and do not necessarily reflect the U.S. Bureau of Labor Statistics policy.

APPENDIX A

From the discussion in Section 2.2 we know that $n\hat{p}_A$ has a binomial distribution $\text{Bin}(n, p_A)$, hence, for $\hat{p}_A = 0, 1/n, 2/n, \dots, 1$,

$$f(\hat{p}_A | p_A) = \frac{\Gamma(n+1)}{\Gamma(n+2)\Gamma(n\hat{p}_A+1)} \frac{\Gamma(n+2)}{\Gamma(n(1-\hat{p}_A)+1)} \times$$

$$p_A^{(n\hat{p}_A+1)-1} (1-p_A)^{(N(1-\hat{p}_A)+1)-1} = k_{\hat{p}_A}(p_A)/(n+1).$$

For each (fixed) value of \hat{p}_A , the function $k_{\hat{p}_A}(p_A)$ is the pdf of a Beta distribution with parameters $\omega_1 = n\hat{p}_A + 1$ and $\omega_2 = n(1 - \hat{p}_A) + 1$. As both ω_1 and ω_2 will be larger than unity with high probability (at least in most real world situations), it is reasonable to approximate $k_{\hat{p}_A}(p_A)$ with a normal pdf having equivalent mean and variance, which are approximately \hat{p}_A and $\hat{p}_A(1 - \hat{p}_A)/n$ respectively.

Assuming that $p_A \sim N(\mu, \sigma^2)$, it follows that the posterior distribution is

$$h(p_A | \hat{p}_A) = f(\hat{p}_A | p_A) g(p_A) / \int_0^1 f(\hat{p}_A | p_A) g(p_A) dp_A \approx ce^{-\frac{1}{2} \left(\frac{(p_A - \hat{p}_A)^2}{\hat{p}_A(1 - \hat{p}_A)/n} + \frac{(p_A - \mu)^2}{\sigma^2} \right)},$$

where c is the normalizing constant.

Under the "empirical Bayes" assumption that $\mu = \hat{p}_A$ and $\sigma^2 = \hat{p}_A(1 - \hat{p}_A)/n$ we have

$$h(p_A | \hat{p}_A) \approx \frac{1}{\sqrt{2\pi} \sqrt{\hat{p}_A(1 - \hat{p}_A)/2n}} e^{-\frac{1}{2} \left(\frac{(p_A - \hat{p}_A)^2}{\hat{p}_A(1 - \hat{p}_A)/2n} \right)}.$$

If we drop the specific assumption regarding σ^2 , and let $\psi = (\hat{p}_A(1 - \hat{p}_A)/n)/\sigma^2$ then $[p_A | \hat{p}_A] \sim N(\hat{p}_A, \hat{p}_A(1 - \hat{p}_A)/(1 + \psi)n)$.

APPENDIX B

Result: Assume W is distributed $N(0, c^2)$ and, conditional on $W = w$, the random variable T is distributed as a non-central t with ν degrees of freedom and non-centrality parameter w . Then, the unconditional distribution of $T/\sqrt{c^2 + 1}$ is central t with ν degrees of freedom.

Proof: First notice that T can be written as $T = (X + W)/\sqrt{S^2/\nu}$, where X is distributed as $N(0, 1)$, S^2 is distributed as χ^2_ν , and X , W , and S^2 are mutually independent. Therefore, $X' = (X + W)/\sqrt{1 + c^2}$ is distributed as $N(0, 1)$. As X' and S^2 are independent, it follows by definition that $T' = T/\sqrt{1 + c^2} = X'/\sqrt{S^2/\nu}$ is distributed as t_ν .

REFERENCES

- DORFMAN, A., and VALLIANT, R. (1993). Quantile variance estimators in complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 866-871.
- JOHNSON, E.G., and RUST, K.F. (1993). Effective Degrees of Freedom for Variance Estimates from a Complex Sample Survey. Paper presented at the 1993 Joint Statistical Meetings, San Francisco.
- KOTT, P.S. (1994). A hypothesis test of linear regression coefficients with survey data. *Survey Methodology*, 20, 159-164.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SATTERTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

On Regression Estimation of Finite Population Means

GIORGIO E. MONTANARI¹

ABSTRACT

This paper examines the main properties of the generalized regression estimator of a finite population mean and those of the regression estimator obtained from the optimal difference estimator. Given that the latter can be more efficient than the former, conditions allowing this to happen are established, and a criterion for choosing between the two types of regression estimators follows. A simulation study illustrates their finite sample performances.

KEY WORDS: Generalized regression estimator; Difference estimator; Auxiliary information.

1. INTRODUCTION

Regression estimation is an effective technique for estimating survey variable finite population means or totals when the population means or totals of a set of auxiliary variables are known. The problem can be stated as follows. Consider a finite population $\mathcal{P} = \{a_1, a_2, \dots, a_N\}$ consisting of N units labelled $1, 2, \dots, N$. Let Y_i be the value of unit a_i of a survey variable y whose population mean $\bar{Y} = \sum_1^N Y_i/N$ has to be estimated by means of a sample drawn from \mathcal{P} . To this end let us suppose that the population mean $\bar{X} = \sum_1^N \mathbf{x}_i/N$ of a q -dimensional auxiliary variable vector, having $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iq})'$ as its value for unit a_i , is known, for example from administrative registers or a census. The entries of \mathbf{x}_i can be quantitative as well as indicator variables denoting the membership of the unit to given subpopulations. Let s be the set of sample unit labels obtained from a sampling design having first order inclusion probabilities $\pi_i, i = 1, 2, \dots, N$, strictly positive. Then, a regression estimator can be written as follows

$$\hat{\bar{Y}}_r = \hat{\bar{Y}} + (\bar{X} - \hat{\bar{X}})' \hat{\boldsymbol{\beta}}, \quad (1)$$

where $\hat{\bar{Y}} = \sum_{i \in s} Y_i/N\pi_i$ and $\hat{\bar{X}} = \sum_{i \in s} \mathbf{x}_i/N\pi_i$ are the Horvitz-Thompson unbiased estimators of \bar{Y} and \bar{X} , respectively, and $\hat{\boldsymbol{\beta}}$ is a vector of regression coefficients, given by some function of sample data $\{(Y_i, \mathbf{x}_i'), i \in s\}$. Briefly, $\hat{\bar{Y}}_r$ is obtained by adding to the unbiased estimator $\hat{\bar{Y}}$ terms proportional to the difference between the true means of the auxiliary variables, $\bar{X}_k = \sum_1^N x_{ki}/N, k = 1, 2, \dots, q$, and the corresponding estimates $\hat{\bar{X}}_k = \sum_{i \in s} x_{ki}/N\pi_i$.

This paper discusses the two chief methods of constructing the vector $\hat{\boldsymbol{\beta}}$ and the properties of the corresponding regression estimators. A criterion based on a first order approximation analysis is then given for selecting one of the two alternatives. Finally, the results of two empirical studies, carried out to explore the finite

sample performances of the examined estimators, are reported. All unsubscripted expectations and variances are taken with respect to a sample design. When calculations are made with respect to a model, a subscript m will be used.

2. MAIN PROPERTIES OF THE REGRESSION ESTIMATOR

Mild restrictions on the second order inclusion probabilities of the sampling design and on the limiting population moments of Y_i and \mathbf{x}_i are sufficient to ensure that the estimator $\hat{\bar{Y}}_r$ can be approximated by the difference estimator

$$\tilde{\bar{Y}}_r = \hat{\bar{Y}} + (\bar{X} - \hat{\bar{X}})' \tilde{\boldsymbol{\beta}}, \quad (2)$$

where $\tilde{\boldsymbol{\beta}}$ is the limit in probability of the vector $\hat{\boldsymbol{\beta}}$, when both the sample size and the population size go to infinity, and the limit is defined as in Isaki and Fuller (1992): Wright (1983); Montanari (1987). Then, the large sample performance of the regression estimator can be studied by means of its linear approximation (2). As a consequence, the regression estimator $\hat{\bar{Y}}_r$ is approximately unbiased, because $\tilde{\bar{Y}}_r$ is unbiased. The sampling variance of $\hat{\bar{Y}}_r$ can be approximated by that of $\tilde{\bar{Y}}_r$, given by

$$V(\tilde{\bar{Y}}_r) = V(\hat{\bar{Y}}) + \tilde{\boldsymbol{\beta}}' V(\hat{\bar{X}}) \tilde{\boldsymbol{\beta}} - 2 \tilde{\boldsymbol{\beta}}' C(\hat{\bar{X}}, \hat{\bar{Y}}), \quad (3)$$

where $V(\hat{\bar{Y}})$ is the variance of $\hat{\bar{Y}}$, $V(\hat{\bar{X}})$ is the $q \times q$ dimensional variance matrix of $\hat{\bar{X}}$, and $C(\hat{\bar{X}}, \hat{\bar{Y}})$ is the q dimensional covariance vector between $\hat{\bar{X}}$ and $\hat{\bar{Y}}$. Since $\tilde{\bar{Y}}_r$ can be rewritten

$$\tilde{\bar{Y}}_r = \bar{X}' \tilde{\boldsymbol{\beta}} + \sum_{i \in s} \frac{U_i}{N\pi_i},$$

¹ Giorgio E. Montanari, Dipartimento di Scienze Statistiche, Università di Perugia, Via A. Pascoli - 06100 Perugia, Italy.

where $U_i = Y_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}$, then

$$V(\tilde{Y}_r) = \sum_{i=1}^N U_i^2 \frac{1 - \pi_i}{N^2 \pi_i} + \sum_{i=1}^N \sum_{j \neq i}^N U_i U_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j}.$$

An approximately unbiased estimator of $V(\tilde{Y}_r)$ is given by the Horvitz-Thompson formula

$$\hat{V}(\tilde{Y}_r) = \sum_{i \in S} \hat{U}_i^2 \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{i \in S} \sum_{j \neq i} \hat{U}_i \hat{U}_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j},$$

where $\hat{U}_i = Y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$. Alternatively, when the sample size is fixed, the Yates-Grundy variance estimator is available, *i.e.*

$$\hat{V}(\tilde{Y}_r) = \sum_{i \in S} \sum_{j > i} \frac{(\pi_i \pi_j - \pi_{ij})}{N^2 \pi_{ij}} \left(\frac{\hat{U}_i}{\pi_i} - \frac{\hat{U}_j}{\pi_j} \right)^2.$$

Henceforth $V(\tilde{Y}_r)$ will be called asymptotic variance of \tilde{Y}_r .

3. THE GENERALIZED REGRESSION ESTIMATOR

Two methods are generally used for constructing the vector $\hat{\boldsymbol{\beta}}$. The first one has been developed within the framework of the model assisted approach to survey sampling inference, as it is described in Särndal, Swensson and Wretman (1992; sec. 6.4) and Estevao, Hidioglou and Särndal (1995). Letting Y_i be either a random variable or an observation of it, consider the following linear regression superpopulation model

$$\begin{cases} E_m(Y_i) = \mathbf{x}_i' \boldsymbol{\beta}, & i = 1, 2, \dots, N, \\ V_m(Y_i) = \sigma^2 v_i, \\ C_m(Y_i, Y_j) = 0, & i \neq j, \end{cases} \quad (4)$$

where E_m , V_m and C_m denote expected value, variance and covariance with respect to the model; $\boldsymbol{\beta}$ and σ^2 are unknown model parameters; v_i is a known function of \mathbf{x}_i . The vector

$$\hat{\boldsymbol{\beta}}_1 = \left[\sum_{i=1}^N \frac{\mathbf{x}_i \mathbf{x}_i'}{v_i} \right]^{-1} \sum_{i=1}^N \frac{\mathbf{x}_i Y_i}{v_i}$$

is the census least squares estimator of $\boldsymbol{\beta}$. Under general conditions, such as those quoted in the referenced papers,

$$\hat{\boldsymbol{\beta}}_1 = \left[\sum_{i \in S} \frac{\mathbf{x}_i \mathbf{x}_i'}{\pi_i v_i} \right]^{-1} \sum_{i \in S} \frac{\mathbf{x}_i Y_i}{\pi_i v_i}, \quad (5)$$

is a consistent estimator of $\tilde{\boldsymbol{\beta}}_1$ and when replaced in (1) gives the generalized regression (GREG) estimator

$$\hat{\tilde{Y}}_{r1} = \hat{\tilde{Y}} + (\bar{X} - \hat{\tilde{X}})' \hat{\boldsymbol{\beta}}_1. \quad (6)$$

In addition to those stated in section 2, this estimator has the following properties: (i) the means of the auxiliary variables estimated through GREG equal the corresponding known population means, *i.e.* $\hat{\tilde{X}}_{r1} = \bar{X}$; (ii) the model expected value of the asymptotic sampling variance, *i.e.* $E_m V(\hat{\tilde{Y}}_{r1})$, is a minimum among all asymptotically design-unbiased estimators of \bar{Y} (Wright 1983). Consequently, if the model is well specified, no other asymptotically unbiased estimator exists that is on the average (with respect to the model) more efficient than $\hat{\tilde{Y}}_{r1}$.

Well known estimators currently used in practice, such as the ratio and post-stratified estimator, belong to the class of GREG estimators. Furthermore, such a class has recently been extended by means of the calibration technique (Deville and Särndal 1992) to better control the variability of the final observation weights.

4. THE OPTIMAL ESTIMATOR

For constructing an alternative regression estimator based on the same auxiliary variable \mathbf{x}_i , a second approach considers the vector $\tilde{\boldsymbol{\beta}}$ that minimizes the asymptotic variance (3) of the difference estimator (2). Assuming $V(\hat{\tilde{X}})$ non singular, *i.e.* there are no linear combinations of the entries of $\hat{\tilde{X}}$ with a zero sampling variance, the minimum variance vector is given by

$$\tilde{\boldsymbol{\beta}}_2 = [V(\hat{\tilde{X}})]^{-1} C(\hat{\tilde{X}}, \hat{\tilde{Y}}).$$

Now, consider the unbiased estimators $\hat{V}(\hat{\tilde{X}})$ and $\hat{C}(\hat{\tilde{X}}, \hat{\tilde{Y}})$ of $V(\hat{\tilde{X}})$ and $C(\hat{\tilde{X}}, \hat{\tilde{Y}})$, respectively, that exist provided that the second order inclusion probabilities of the sample design are all positive. They are given by the Horvitz-Thompson formula or the Yates-Grundy formula when applicable. For example, using the former we have the estimated covariance vector

$$\hat{C}(\hat{\tilde{X}}, \hat{\tilde{Y}}) = \sum_{i \in S} \mathbf{x}_i Y_i \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{i \in S} \sum_{j \neq i} \mathbf{x}_i Y_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j \pi_{ij}}.$$

Using $\hat{V}(\hat{\tilde{X}})$ and $\hat{C}(\hat{\tilde{X}}, \hat{\tilde{Y}})$ we get the alternative regression estimator

$$\hat{\tilde{Y}}_{r2} = \hat{\tilde{Y}} + (\bar{X} - \hat{\tilde{X}})' \hat{\boldsymbol{\beta}}_2,$$

where $\hat{\boldsymbol{\beta}}_2 = [\hat{V}(\hat{\tilde{X}})]^{-1} \hat{C}(\hat{\tilde{X}}, \hat{\tilde{Y}})$. It was studied by Montanari (1987) and called by Rao (1994) the optimal estimator. When $V(\hat{\tilde{X}})$ is singular and its rank is $q' < q$, to

define the optimal estimator it is understood that one or more entries of \mathbf{x}_i , hence of $\hat{\mathbf{X}}$ have to be dropped in such a way as to obtain a $q' \times q'$ non singular variance matrix.

Using the expression for $\tilde{\beta}_2$, the asymptotic variance of \tilde{Y}_{r2} simplifies to

$$V(\tilde{Y}_{r2}) = V(\tilde{Y}) - C(\hat{\mathbf{X}}, \hat{\mathbf{Y}})' [V(\hat{\mathbf{X}})]^{-1} C(\hat{\mathbf{X}}, \hat{\mathbf{Y}}). \quad (7)$$

The properties of the optimal estimator are: (i) asymptotically, the efficiency of \tilde{Y}_{r2} is not inferior to that of \tilde{Y}_{r1} , i.e., $V(\tilde{Y}_{r2}) \leq V(\tilde{Y}_{r1})$; (ii) the means of the auxiliary variables estimated through the optimal estimator equal the corresponding known population means, i.e. $\hat{\mathbf{X}}_{r2} = \bar{\mathbf{X}}$. As for the case of the GREG estimator, when there is more than one survey variable, the optimal estimator \tilde{Y}_{r2} can be expressed as a simple weighted estimator with the same weights applying to all variables of interest. For example, using the Horvitz-Thompson formula for variance and covariance estimators, we can write $\tilde{Y}_{r2} = \sum_{i \in S} Y_i w_i$ where

$$w_i = \frac{1}{\pi_i} + (\bar{\mathbf{X}} - \hat{\mathbf{X}})' [\hat{V}(\hat{\mathbf{X}})]^{-1} \left(\mathbf{x}_i \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{j \neq i} \mathbf{x}_j \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j \pi_{ij}} \right).$$

A similar result can be achieved with the Yates-Grundy formula.

Note that the asymptotic optimality of \tilde{Y}_{r2} is a strictly design based property, achieved conditionally on the realized finite population (hence, within the fixed population approach to the finite population inference). On the contrary, the asymptotic optimality of \tilde{Y}_{r1} requires the model to be true, and concerns the average asymptotic variance over the finite populations that can be generated under the model.

Because of these results, \tilde{Y}_{r2} would seem preferable to \tilde{Y}_{r1} . However, $\hat{\beta}_1$ is a function of population total estimators, and $\hat{\beta}_2$ is a function of variance and covariance estimators. As a consequence, the former is more vulnerable to model misspecification, and the latter is more vulnerable to sampling fluctuations. In a finite size sample, \tilde{Y}_{r2} is generally less stable and more complex to compute and its variance can be greater than that of \tilde{Y}_{r1} ; see Casady and Valliant (1993). However, if an adequate number, g , of degrees of freedom are available for estimating β_2 , the instability problem of \tilde{Y}_{r2} can be overcome. For example, for standard complex sampling designs having with-replacement sampling at the first stage, g can be roughly taken as the number of sample clusters minus the number of strata (Lehtonen and Pahkinen 1995; p. 181; see Eltinge and Jang 1996, for more elaboration on this topic). A stable $\hat{\beta}_2$ can be expected when g is large enough relative to the dimension q of the auxiliary variable \mathbf{x}_i . Since with

modern computers the computation of \hat{Y}_{r2} is less problematic, it becomes interesting to develop a criterion for recognizing when such an estimator is truly advantageous.

5. A CRITERION FOR CHOOSING BETWEEN \tilde{Y}_{r1} AND \tilde{Y}_{r2}

Consider the following theorem:

Theorem: Let $V(\tilde{Y}_r)$ and $V(\tilde{Y}_{r2})$ be the asymptotic variances of the general regression estimator \tilde{Y}_r and the optimal estimator \tilde{Y}_{r2} , respectively. Then

$$V(\tilde{Y}_r) - V(\tilde{Y}_{r2}) = C(\hat{\mathbf{X}}, \tilde{\mathbf{Y}})' [V(\hat{\mathbf{X}})]^{-1} C(\hat{\mathbf{X}}, \tilde{\mathbf{Y}}). \quad (8)$$

Proof: Using (3) and (7), the difference in variances is

$$V(\tilde{Y}_r) - V(\tilde{Y}_{r2}) = \tilde{\beta}' V(\hat{\mathbf{X}}) \tilde{\beta} - 2\tilde{\beta}' C(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) + C(\hat{\mathbf{X}}, \hat{\mathbf{Y}})' [V\hat{\mathbf{X}}]^{-1} C(\hat{\mathbf{X}}, \hat{\mathbf{Y}}).$$

Since $\tilde{\beta}_2 = [V(\hat{\mathbf{X}})]^{-1} C(\hat{\mathbf{X}}, \hat{\mathbf{Y}})$ and $\tilde{\beta}' C(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) = \tilde{\beta}' V(\hat{\mathbf{X}}) \tilde{\beta}_2$ we have

$$V(\tilde{Y}_r) - V(\tilde{Y}_{r2}) = (\tilde{\beta} - \tilde{\beta}_2)' V(\hat{\mathbf{X}}) (\tilde{\beta} - \tilde{\beta}_2).$$

But, $C(\hat{\mathbf{X}}, \tilde{\mathbf{Y}}) = C(\hat{\mathbf{X}}, \hat{\mathbf{Y}}) - V(\hat{\mathbf{X}}) \tilde{\beta} = V(\hat{\mathbf{X}}) (\tilde{\beta}_2 - \tilde{\beta})$ and (8) follows.

Note that the right hand side of (8) is a positive definite quadratic form and it is equal to zero if and only if $C(\hat{\mathbf{X}}, \tilde{\mathbf{Y}}) = \mathbf{0}$. Therefore, the smaller the absolute values of the entries of $C(\hat{\mathbf{X}}, \tilde{\mathbf{Y}})$ are, the smaller the difference $V(\tilde{Y}_r) - V(\tilde{Y}_{r2})$ is. The main conclusion the theorem provides us is that an efficient use of any known auxiliary variable population mean requires us to adopt estimators that are uncorrelated with the auxiliary variable mean estimator.

Applying the theorem to the GREG estimator, let us consider the k -th entry of $C(\hat{\mathbf{X}}, \tilde{\mathbf{Y}}_{r1})$ that can be written

$$C(\hat{X}_k, \tilde{Y}_{r1}) = \sum_{i=1}^N U_i x_{ki} \frac{1 - \pi_i}{N^2 \pi_i} + \sum_{i=1}^N \sum_{j \neq i}^N U_i x_{kj} \frac{\pi_{ij} - \pi_i \pi_j}{N^2 \pi_i \pi_j},$$

where $U_i = Y_i - \mathbf{x}_i' \tilde{\beta}_1$. If the superpopulation model (4) is well specified, it follows that $E_m(U_i) = 0$, for all i , and $E_m[C(\hat{X}_k, \tilde{Y}_{r1})] = 0$. Therefore, $C(\hat{X}_k, \tilde{Y}_{r1})$ must be approximately zero for all $k = 1, 2, \dots, q$, being proportional to a weighted average of N uncorrelated random variables with expected values zero. Consequently the difference $V(\tilde{Y}_{r1}) - V(\tilde{Y}_{r2})$ must be negligible. The result suggests

using the more practical \hat{Y}_{r1} . The conclusion is that the estimator \hat{Y}_{r2} can achieve substantial gains in efficiency compared to \hat{Y}_{r1} if the superpopulation model upon which the latter is based is not good enough. This can happen because of the specification of the linear superpopulation model is being confined to regressors with a known population mean.

Since the following quantity

$$\lambda(\hat{Y}_{r1}, \hat{Y}_{r2}) = C(\hat{X}, \tilde{Y}_{r1})' [V(\hat{X})]^{-1} C(\hat{X}, \tilde{Y}_{r1}) / V(\tilde{Y}_{r1})$$

gives the asymptotic relative gain in efficiency that can be achieved with \hat{Y}_{r2} compared to \hat{Y}_{r1} , we propose it as an indicator of a model inadequacy for extracting all information from the sample. When $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ is greater than 10% or 15%, say, the optimal estimator should be adopted. Provided that the second order inclusion probabilities are all positive, under general conditions $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ can be consistently estimated from sample data. Then, the information offered by the estimate $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$ can be used for shifting from \hat{Y}_{r1} to \hat{Y}_{r2} in the next repetition of a periodic survey, or, as we suggest in section 6, within the same survey, choosing between \hat{Y}_{r1} and \hat{Y}_{r2} at the estimation stage.

This section concludes with a few examples.

Example 1. Consider a simple random sample of n units and the linear regression model through the origin $E_m(Y_i) = x_i\beta$, $V_m(Y_i) = \sigma^2 x_i$, $C_m(Y_i, Y_j) = 0$, $i \neq j$, assuming \bar{X} known. In this case the GREG is the ratio estimator of the mean, i.e., $\hat{Y}_{r1} = \bar{X}\bar{y}/\bar{x}$, where \bar{y} and \bar{x} are the sample means of y and x , respectively. The linear approximation is $\tilde{Y}_{r1} = \bar{X}R + \sum_{i \in S} U_i/n$, where $U_i = Y_i - Rx_i$ and $R = \bar{y}/\bar{x}$. Then, the covariance of \bar{x} and \tilde{Y}_{r1} is

$$C(\bar{x}, \tilde{Y}_{r1}) = \frac{N-n}{Nn} S_x^2 \left[\frac{S_{xy}}{S_x^2} - R \right], \quad (9)$$

where S_{xy} is the population covariance between y and x and S_x^2 is the population variance of x . If the model is well specified, then $S_{yx}/S_x^2 \approx R$ and expression (9) must be approximately zero. Otherwise, the greater the absolute value of an intercept in a census linear regression of y on x , the more \hat{Y}_{r2} is asymptotically efficient than \hat{Y}_{r1} . The result is not new (for example, see Cochran 1977; sec. 7.5), but it is achieved within the framework of a general theory. Note that $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2}) = [S_{xy}/S_x^2 - R]^2 S_x^2 / S_u^2$, where S_u^2 is the population variance of U_i , is a constant with respect to the sample size. When $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ is not negligible, \hat{Y}_{r2} should be chosen as regression estimator, or, alternatively, an intercept plugged into the model in order to use the corresponding GREG estimator \hat{Y}_{r1} . However, for simple random sampling both solutions give the same estimator, i.e., $\hat{Y}_{r1} = \hat{Y}_{r2}$, but in general they are different, even for self-weighting designs.

Example 2. Consider a stratified random sample and the linear homoscedastic regression model $E_m(Y_i) = \alpha + x_i\beta$, $V_m(Y_i) = \sigma^2$, $C_m(Y_i, Y_j) = 0$, $i \neq j$. Assume that \bar{X} is known and that individual x_i 's are known only for sample units and not for the nonsampled units. Now, the auxiliary information is given by $\mathbf{x}_i = (1, x_i)'$ and the corresponding GREG estimator can be written $\hat{Y}_{r1} = \hat{Y} + (\bar{X} - \hat{X})\hat{\beta}_1$, where

$$\hat{\beta}_1 = \frac{(\sum_{i \in S} Y_i x_i / N\pi_i) - \hat{X}\hat{Y}}{(\sum_{i \in S} x_i^2 / N\pi_i) - \hat{X}^2},$$

and where the estimated α cancels out. Because $\tilde{\beta}_1 = S_{yx}/S_x^2$ and $U_i = Y_i - \bar{Y} - \tilde{\beta}_1(x_i - \bar{X})$, we have

$$C(\hat{X}, \tilde{Y}_{r1}) = \sum_{h=1}^H \frac{N_h(N_h - n_h)}{N^2 n_h} S_{hx}^2 (\tilde{\beta}_{h1} - \tilde{\beta}_1), \quad (10)$$

where the subindex h denotes stratum quantities and $\tilde{\beta}_{h1} = S_{hxy}/S_{hx}^2$. The right hand side of (10) is a function of the differences between each within-stratum regression coefficient and the coefficient for the whole population. If the model is well specified, the differences $\tilde{\beta}_{h1} - \tilde{\beta}_1$ must be negligible. Otherwise, $C(\hat{X}, \tilde{Y}_{r1})$ can take non negligible absolute values and, since only \bar{X} is known, the estimator \hat{Y}_{r2} appears to extract better all the information from the sample value of \hat{X} .

It is interesting to note that when the allocation of the sample is proportional, i.e., $n_h \propto N_h$, ignoring terms of order $1/N_h$ relative to unity, \hat{Y}_{r2} is equal to the GREG estimator based on the auxiliary variable $\mathbf{x}_i = (d_{1i}, d_{2i}, \dots, d_{Hi}, x_i)'$ and $v_i = 1$, where d_{hi} is an indicator variable of the membership of unit i to stratum $h = 1, 2, \dots, H$. This model fits different regression lines with a common slope within the strata.

Example 3. Consider a complex sampling design and suppose that the population can be partitioned into H post-strata of known sizes. Assume the superpopulation model $E_m(Y_i) = \beta_{h(i)}$, $V_m(Y_i) = \sigma^2$, and $C_m(Y_i, Y_j) = 0$, $i \neq j$, where the subindex $h(i)$ denotes the post-stratum to which the i -th unit belongs. Denoting by d_{hi} the indicator variable of the i -th unit membership to post-stratum h , and with \bar{D}_h its known population mean, putting $\mathbf{x}_i = (d_{1i}, d_{2i}, \dots, d_{Hi})'$ and $v_i = 1$, in (5), we get the post-stratified estimator, $\hat{Y}_{r1} = \sum_1^H \bar{D}_h \hat{Z}_h / \hat{\bar{D}}_h$, where \hat{Z}_h and $\hat{\bar{D}}_h$ are the Horvitz-Thompson mean estimators of the variables $z_{hi} = Y_i d_{hi}$ and d_{hi} , respectively. The linear approximation is $\tilde{Y}_r = \bar{Y} + (\bar{X} - \hat{X})' \tilde{\beta}_1$, where $\tilde{\beta}_1 = (R_1, R_2, \dots, R_H)'$, $R_h = \bar{Z}_h / \bar{D}_h$ (i.e., the mean value of y in the h -th post-stratum), and $\hat{X} = (\hat{\bar{D}}_1, \hat{\bar{D}}_2, \dots, \hat{\bar{D}}_H)'$. Since $U_i = Y_i - \sum_1^H R_h d_{hi}$, the covariance of $\hat{\bar{D}}_h$ and \tilde{Y}_{r1} is

$$C(\tilde{Y}_{r1}, \hat{\bar{D}}_h) = C(\hat{Y}, \hat{\bar{D}}_h) - \sum_{j=1}^H R_j C(\hat{\bar{D}}_j, \hat{\bar{D}}_h). \quad (11)$$

Under the superpopulation model upon which $\hat{\bar{Y}}_{r1}$ is based on, we have $E_m[C(\hat{\bar{Y}}_{r1}, \hat{\bar{D}}_h)] = 0$ and a negligible value of $C(\hat{\bar{Y}}_{r1}, \hat{\bar{D}}_h)$ is expected for all h . It can be easily seen that for simple random sampling, formula (11) is identically zero. But in complex sampling schemes such covariances might take non negligible values, for example, when in a multistage sampling scheme a linear regression of the primary unit totals of z_{hi} on the totals of d_{hi} yields a non negligible intercept for some h . See Casady and Valliant (1993) for a case study.

6. EMPIRICAL STUDIES

The above analysis is based on first order approximations. In the following empirical studies the finite sample performances of $\hat{\bar{Y}}_{r1}$ and $\hat{\bar{Y}}_{r2}$ will be explored within the framework of example 2.

6.1 The First Empirical Study

In this first empirical study we consider a population of infinite size subdivided into two strata of equal weights and a proportional stratified random sampling design to estimate the mean of a survey variable y . To this end, let us suppose that there exists a scalar variable x that was not available for stratification but with a known population mean \bar{X} and unknown stratum means (*i.e.*, the x values are not available for nonsampled units).

Since only the population mean of x is assumed known, a reasonable superpopulation model that can be assumed to identify a GREG estimator is the linear regression one, with homoscedastic errors, *i.e.*, $E_m(Y_i) = \alpha + x_i\beta$, $V_m(Y_i) = \sigma^2$, $C_m(Y_i, Y_j) = 0$, $i \neq j$. The auxiliary variable plugged into (5) is $\mathbf{x}_i = (1, x_i)'$ and the corresponding GREG estimator can be written

$$\hat{\bar{Y}}_{r1} = \bar{y} + (\bar{X} - \bar{x})s_{yx}/s_x^2,$$

where \bar{y} and \bar{x} are the sample means of y and x , s_{yx} is the sample covariance between y and x , and s_x^2 is the sample variance of x . The linear approximation is

$$\bar{\bar{Y}}_{r1} = \bar{y} + (\bar{X} - \bar{x})S_{yx}/S_x^2,$$

where S_{yx} and S_x^2 are the population analogues of s_{yx} and s_x^2 .

Dropping the first component of $\mathbf{x}_i = (1, x_i)'$, whose mean is estimated without error, the optimal estimator based on the same auxiliary variable is given by

$$\hat{\bar{Y}}_{r2} = \bar{y} + (\bar{X} - \bar{x})\hat{C}(\bar{y}, \bar{x})/\hat{V}(\bar{x}),$$

where \bar{X} is the population mean of x , $\hat{C}(\bar{y}, \bar{x})$ and $\hat{V}(\bar{x})$ are the standard unbiased estimators of the covariance

between \bar{y} and \bar{x} and the variance of \bar{x} , respectively. The corresponding linear approximation is

$$\bar{\bar{Y}}_{r2} = \bar{y} + (\bar{X} - \bar{x})C(\bar{y}, \bar{x})/V(\bar{x}),$$

where $C(\bar{y}, \bar{x})$ and $V(\bar{x})$ are the true covariance and variance.

In this case, the expression of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$ simplifies to

$$\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = \frac{\sum_1^2 S_{hx}^2}{\sum_1^2 S_{hu}^2} \left(\frac{\sum_1^2 S_{hxy}}{\sum_1^2 S_{hx}^2} - \frac{S_{yx}}{S_x^2} \right),$$

and it can be estimated replacing the population variances and covariances with the sample analogues.

Four simulations were performed. In the first two, the sample values of x were drawn from a uniform distribution on [30–70] in the first stratum and [50–90] in the second one. The sample values of y , given x , were drawn from a normal distribution with expected values $1.26x$ in the first stratum and $0.82x$ in the second. The conditional variance was $8x$ in both strata in the first simulation and $3x$ in the second one. In the third and fourth simulation, the sample values of x were drawn from a linearly transformed gamma random variable with parameters chosen to achieve the first two simulation stratum means and variances for x and y and an asymmetry index for x (given by the ratio between the third central moment and the third power of the standard deviation) equal to 2.5. This allows studying the effects of a strong asymmetry in the marginal distributions of y and x .

The populations were constructed to have $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = 8.1\%$ when $V(Y|x) = 8x$, and $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) = 18.6\%$, when $V(Y|x) = 3x$. Note that the GREG estimator based on the true model is the separate ratio estimator; however, its use would require the knowledge of the stratum means of x , but they are assumed unknown.

In each simulation we drew 10,000 samples of size 20 (ten units per stratum), and 5,000 of size 40 (twenty units per stratum). For each sample we computed the values of the Horvitz-Thompson estimator $\hat{\bar{Y}} = \bar{y}$, and of $\hat{\bar{Y}}_{r1}$, $\hat{\bar{Y}}_{r2}$, $\hat{\bar{Y}}_{r1}$, $\hat{\bar{Y}}_{r2}$, and $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. We also computed an estimator $\hat{\bar{Y}}_{r3}$, defined to take the value of $\hat{\bar{Y}}_{r1}$, when $\hat{\lambda}(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2}) \leq 8\%$, and the value of $\hat{\bar{Y}}_{r2}$ otherwise. So, $\hat{\bar{Y}}_{r3}$ is a sample dependent type estimator, constructed choosing between $\hat{\bar{Y}}_{r1}$ and $\hat{\bar{Y}}_{r2}$ according to the estimated value of $\lambda(\hat{\bar{Y}}_{r1}, \hat{\bar{Y}}_{r2})$. Here, 8% is an arbitrarily chosen threshold, over which shifting from $\hat{\bar{Y}}_{r1}$ to $\hat{\bar{Y}}_{r2}$ is thought to be convenient.

Table 1 reports for each simulation the empirical results achieved with reference to the percent relative bias of estimators (RB) and the mean squared error (MSE), in the latter case having set that of the Horvitz-Thompson estimators equal to 100 by multiplying the MSE values by $100/\text{MSE}(\bar{y})$. As we can see, the biases are all negligible

(the biggest absolute value is less than 0.6% and all biases are less than 10% of the corresponding standard errors) and contribute to the MSE in a negligible manner. The MSE reduction percentages that can be achieved shifting from \tilde{Y}_{r1} to \tilde{Y}_{r2} are approximately equal to the fixed in advance values of $\lambda(\tilde{Y}_{r1}, \tilde{Y}_{r2})$, i.e., 8.1% and 18.6%. The effective MSE values of \tilde{Y}_{r1} and \tilde{Y}_{r2} are greater than the corresponding asymptotic values, in particular when the population is asymmetric and the estimator is the optimal one. For example, in the third simulation, when $n = 20$, the MSE of \tilde{Y}_{r1} shows a 5.1% relative increase compared to that of \tilde{Y}_{r1} , while the corresponding value for \tilde{Y}_{r2} is 10.7%. Doubling the sample size, those relative values decrease to 2.8% and 3.6%, respectively. As we observed in example 2, when the sample allocation is proportional, \tilde{Y}_{r2} is equal to the GREG estimator based on a homoscedastic linear model that fits two parallel regression lines in the two strata. So, the greater loss in efficiency percentage of \tilde{Y}_{r2} with respect to its asymptotic variance can be explained by the added parameter to be estimated in the model.

The performance of $\hat{\tilde{Y}}_{r3}$ is also interesting; this estimator is approximately unbiased and its MSE is lower than that of $\hat{\tilde{Y}}_{r1}$ the more often $\hat{\tilde{Y}}_{r2}$ is selected. Table 1 reports for each simulation the percentages of samples for which $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2}) > 8\%$ and $\hat{\tilde{Y}}_{r2}$ was selected instead of $\hat{\tilde{Y}}_{r1}$. The higher is the theoretic value of $\lambda(\tilde{Y}_{r1}, \tilde{Y}_{r2})$, the more often $\hat{\tilde{Y}}_{r2}$ is chosen over $\hat{\tilde{Y}}_{r1}$.

Obviously, the performance of $\hat{\tilde{Y}}_{r3}$ depends on the sampling distribution of the sample statistics $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$. Table 2 reports the means, the standard deviations, and some quantiles of the empirical distributions of $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$ for the gamma populations, which are the more problematic ones. As it can be seen, the distributions of $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$ were in all cases positively skewed and highly variable. This means that larger sample sizes than those considered here are needed to get reliable estimates of $\lambda(\tilde{Y}_{r1}, \tilde{Y}_{r2})$. Clearly, the less the variance of $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$, the higher is the gain in efficiency of $\hat{\tilde{Y}}_{r3}$ over $\hat{\tilde{Y}}_{r1}$ when the true value of $\lambda(\tilde{Y}_{r1}, \tilde{Y}_{r2})$ is over the threshold for $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$ chosen to shift from $\hat{\tilde{Y}}_{r1}$ to $\hat{\tilde{Y}}_{r2}$.

Table 1
Empirical percent relative bias (RB) and Mean Squared Error (MSE) of \bar{y} , \tilde{Y}_{r1} , \tilde{Y}_{r2} , $\hat{\tilde{Y}}_{r1}$, $\hat{\tilde{Y}}_{r2}$ and $\hat{\tilde{Y}}_{r3}$
and percentage of samples for which $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2}) > 8\%$ in the first empirical study

Uniform populations								
Estimator	$V(Y x) = 8x$				$V(Y x) = 3x$			
	$n = 20$		$n = 40$		$n = 20$		$n = 40$	
	RB (%)	MSE	RB (%)	MSE	RB (%)	MSE	RB (%)	MSE
\bar{y}	-0.06	100.0	-0.08	100.0	0.12	100.0	-0.10	100.0
\tilde{Y}_{r1}	-0.05	83.8	-0.06	84.1	0.10	69.4	-0.05	68.8
\tilde{Y}_{r2}	-0.03	77.3	-0.04	77.7	0.07	56.2	0.01	55.8
$\hat{\tilde{Y}}_{r1}$	0.07	87.7	-0.01	86.2	0.22	73.4	-0.00	70.5
$\hat{\tilde{Y}}_{r2}$	-0.05	82.4	-0.04	80.1	0.05	59.8	-0.00	57.3
$\hat{\tilde{Y}}_{r3}$	-0.06	85.0	-0.05	83.1	0.03	61.0	-0.01	57.9
Freq ($\lambda > 8\%$)	53.5%		53.6%		88.6%		93.5%	
Gamma populations								
Estimator	$V(Y x) = 8x$				$V(Y x) = 3x$			
	$n = 20$		$n = 40$		$n = 20$		$n = 40$	
	RB(%)	MSE	RB(%)	MSE	RB(%)	MSE	RB(%)	MSE
\bar{y}	0.07	100.0	-0.01	100.0	0.02	100.0	-0.03	100.0
\tilde{Y}_{r1}	0.08	84.1	0.02	84.3	0.06	69.8	-0.03	69.9
\tilde{Y}_{r2}	0.09	77.5	0.05	78.1	0.10	57.1	-0.02	56.9
$\hat{\tilde{Y}}_{r1}$	-0.58	88.4	-0.30	86.7	-0.60	75.5	-0.36	72.8
$\hat{\tilde{Y}}_{r2}$	0.03	85.8	0.03	80.9	0.12	63.5	-0.02	59.1
$\hat{\tilde{Y}}_{r3}$	-0.05	87.9	0.07	86.2	0.06	65.4	-0.04	60.8
Freq ($\lambda > 8\%$)	50.6%		50.3%		86.9%		91.7%	

Table 2

Selected characteristics of the empirical distributions of $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$ for gamma populations (first empirical study)

Gamma Populations	Mean	Standard deviation	Median	Quantiles 10% 90%	
$V(Y x) = 8x, n = 20$	10.7	9.8	8.7	1.3	24.9
$V(Y x) = 8x, n = 40$	9.2	6.3	8.3	2.5	19.1
$V(Y x) = 3x, n = 20$	21.6	12.3	19.2	6.9	40.7
$V(Y x) = 3x, n = 40$	19.0	9.5	18.9	9.4	34.2

6.2 The Second Empirical Study

In the second empirical study, we consider a finite population subdivided into eight strata each of size 100, according to an auxiliary variable x whose values are assumed known for each unit of the population. In order to simulate a stratification based on x , the values of x were assigned through the monotonic function of h and i

$$x_{hi} = 4.95 + 5 \sum_{j=1}^{h-1} j + h \cdot i,$$

where hi is the label of the unit $i = 1, 2, \dots, 100$ within the stratum $h = 1, 2, \dots, 8$.

A finite population of y values, given x , was generated using the model

$$Y_{hi} = 20 + 2x_{hi} + 0.06x_{hi}^2 + \epsilon_{hi} \cdot x_{hi},$$

where ϵ_{hi} is a standard normal random variable. The realized values of the mean, standard deviation and asymmetry index of y are 618.2, 676.0, and 1.21, respectively. The correlation between y and x is 0.96.

A proportional stratified random sampling without replacement design was used to select 5,000 samples of size $n = 40$ (five units per stratum) and 2,500 samples of size 80 (ten units per stratum). For each sample we computed the following quantities:

- the unbiased estimator of the population mean \bar{Y} , i.e., \bar{y} ;
- the ratio estimator \hat{Y}_{r11} , based on the model $E_m(Y_{hi}) = \beta x_{hi}$ and $V_m(Y_{hi}) = \sigma^2 x_{hi}$, and obtained from (5) and (6) putting $x_{hi} = x_{hi}$ and $v_{hi} = x_{hi}$;
- the optimal estimator \hat{Y}_{r21} , based on the same auxiliary variable used for \hat{Y}_{r11} ;
- the GREG estimator \hat{Y}_{r12} , based on the model $E_m(Y_{hi}) = \alpha + \beta x_{hi}$ and $V_m(Y_{hi}) = \sigma^2 x_{hi}$, and obtained from (5) and (6) putting $x_{hi} = (1, x_{hi})'$ and $v_{hi} = x_{hi}$;
- the optimal estimator \hat{Y}_{r22} based on the same auxiliary variables used for \hat{Y}_{r12} ;
- the GREG estimator \hat{Y}_{r13} , based on the model $E_m(Y_{hi}) = \alpha + \beta x_{hi} + \gamma x_{hi}^2$ and $V_m(Y_{hi}) = \sigma^2 x_{hi}^2$ (the true model), and obtained from (5) and (6) putting $x_{hi} = (1, x_{hi}, x_{hi}^2)'$ and $v_{hi} = x_{hi}^2$;

- the optimal estimator \hat{Y}_{r23} based on the same auxiliary variables used for \hat{Y}_{r13} ;
- the linear approximations $\tilde{Y}_{r12}, \tilde{Y}_{r13}, \tilde{Y}_{r22}$, and \tilde{Y}_{r23} of $\hat{Y}_{r12}, \hat{Y}_{r13}, \hat{Y}_{r22}$, and \hat{Y}_{r23} , respectively;
- the statistics $\hat{\lambda}(\tilde{Y}_{r1k}, \tilde{Y}_{r2k})$, for $k = 1, 2, 3$;
- the sample dependent estimators \tilde{Y}_{r3k} ($k = 1, 2, 3$) defined to take the value of \tilde{Y}_{r1k} when $\hat{\lambda}(\tilde{Y}_{r1k}, \tilde{Y}_{r2k}) \leq 8\%$, and the value of \tilde{Y}_{r2k} otherwise.

We do not consider separate regression estimation because sample sizes within strata are small. The finite population is such that $\lambda(\tilde{Y}_{r11}, \tilde{Y}_{r21}) = 0.22$, $\lambda(\tilde{Y}_{r12}, \tilde{Y}_{r22}) = 0.16$, and $\lambda(\tilde{Y}_{r13}, \tilde{Y}_{r23}) = 0.00$. Note that because of the sample design considered we have $\tilde{Y}_{r21} = \tilde{Y}_{r22}$ and therefore we omit \tilde{Y}_{r21} .

Table 3 reports the empirical results achieved with reference to the percent relative bias of estimators (RB) and the Mean Squared Error (MSE), in the latter case having set that of the Horvitz-Thompson estimators equal to 100. The results are separated according to the sample size.

Again, the biases are all negligible. The MSE reduction percentage that can be achieved with respect to the sample mean increases with the number of auxiliary variables used. However, as expected \hat{Y}_{r11} and \hat{Y}_{r12} are less efficient than the optimal estimator \hat{Y}_{r22} based on the same auxiliary variables. The statistics $\hat{\lambda}(\tilde{Y}_{r11}, \tilde{Y}_{r21})$ and $\hat{\lambda}(\tilde{Y}_{r12}, \tilde{Y}_{r22})$ take values above the 8% threshold most of the time, especially when the sample size is 80. The sample dependent estimators \hat{Y}_{r31} and \hat{Y}_{r32} are both more efficient than \hat{Y}_{r11} and \hat{Y}_{r12} . The result is due to the inadequacy of the models upon which \hat{Y}_{r11} and \hat{Y}_{r12} are based for extracting all information from the sample. On the other hand, \hat{Y}_{r13} is more efficient than \hat{Y}_{r23} because it is based on the true model. Most of the time the statistic $\hat{\lambda}(\tilde{Y}_{r13}, \tilde{Y}_{r23})$ is below the threshold, especially when the sample size is 80, and the sample dependent estimator \hat{Y}_{r33} is almost as efficient as \hat{Y}_{r13} .

Looking at the linear approximations, first we observe that the MSE's of the GREG estimators \hat{Y}_{r12} and \hat{Y}_{r13} are almost equal to those of \tilde{Y}_{r12} and \tilde{Y}_{r13} in this second study. This is not true for the optimal estimators \hat{Y}_{r22} and \hat{Y}_{r23} . The losses in efficiency with respect to their linear approximations \tilde{Y}_{r22} and \tilde{Y}_{r23} are greater, but they diminish rapidly when the sample size increases. The MSE's of the linear approximations confirm that given a certain amount of auxiliary information, a negligible gain in efficiency can be achieved through the optimal estimator, even with very large samples (compare \tilde{Y}_{r13} with \tilde{Y}_{r23}), when the model upon which the GREG is based holds true. Substantial gains in efficiency can be achieved if the model is not adequate, such as those upon which \hat{Y}_{r11} and \hat{Y}_{r12} are based (compare \tilde{Y}_{r12} with \tilde{Y}_{r22}). Table 4 reports the means, standard deviations and some quantiles of the empirical distributions of $\hat{\lambda}(\tilde{Y}_{r1k}, \tilde{Y}_{r2k})$, $k = 1, 2, 3$.

Table 3

Empirical percent relative bias (RB) and Mean Squared Error (MSE) of estimators and percentage of samples for which $\hat{\lambda}(\hat{\tilde{Y}}_{r1k}, \hat{\tilde{Y}}_{r2k}) > 8\%$ in the second empirical study

Auxiliary used	Estimator	Sample size 40			Sample size 80		
		RB(%)	MSE	($\lambda > 8\%$)	RB(%)	MSE	($\lambda > 8\%$)
none	\bar{y}	0.01	100.0	—	0.01	100.0	—
(x)	$\hat{\tilde{Y}}_{r11}$	-0.01	55.2	82.6%	0.00	54.3	85.0%
(x)	$\hat{\tilde{Y}}_{r31}$	-0.05	48.4	—	-0.02	43.8	—
(1, x)'	$\hat{\tilde{Y}}_{r12}$	-0.01	51.7	72.7%	0.00	50.8	83.2%
(1, x)'	$\hat{\tilde{Y}}_{r22}$	-0.05	47.4	—	-0.01	43.3	—
(1, x)'	$\hat{\tilde{Y}}_{r32}$	-0.05	48.3	—	-0.02	43.8	—
(1, x)'	$\hat{\tilde{Y}}_{r12}$	0.02	51.6	—	0.01	50.7	—
(1, x)'	$\hat{\tilde{Y}}_{r22}$	0.02	44.3	—	0.00	42.3	—
(1, x, x ²)'	$\hat{\tilde{Y}}_{r13}$	-0.01	35.1	28.9%	0.02	33.5	10.5%
(1, x, x ²)'	$\hat{\tilde{Y}}_{r23}$	-0.10	38.0	—	-0.03	34.7	—
(1, x, x ²)'	$\hat{\tilde{Y}}_{r33}$	-0.04	37.0	—	-0.01	33.8	—
(1, x, x ²)'	$\hat{\tilde{Y}}_{r13}$	0.01	34.9	—	0.03	33.5	—
(1, x, x ²)'	$\hat{\tilde{Y}}_{r23}$	0.01	34.7	—	0.03	33.2	—

Table 4

Selected characteristics of the empirical distributions of $\hat{\lambda}(\hat{\tilde{Y}}_{r1k}, \hat{\tilde{Y}}_{r2k})$, $k = 1, 2, 3$ (second empirical study)

Statistics	Sample size 40					Sample size 80				
	Mean	Standard deviation	Median	Quantiles		Mean	Standard deviation	Median	Quantiles	
				10%	90%				10%	90%
$\hat{\lambda}(\hat{\tilde{Y}}_{r11}, \hat{\tilde{Y}}_{r21})$	0.24	0.15	0.23	0.04	0.45	0.23	0.10	0.23	0.07	0.35
$\hat{\lambda}(\hat{\tilde{Y}}_{r12}, \hat{\tilde{Y}}_{r22})$	0.19	0.14	0.17	0.02	0.38	0.18	0.09	0.17	0.04	0.30
$\hat{\lambda}(\hat{\tilde{Y}}_{r13}, \hat{\tilde{Y}}_{r23})$	0.06	0.08	0.03	0.00	0.18	0.03	0.04	0.01	0.00	0.08

7. DISCUSSION

The optimal estimator can be an efficient alternative to the generalized regression estimator based on misspecified superpopulation models when the sample size is large enough. This efficiency can be measured by means of the sample statistic, $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$, that captures the asymptotic relative gain in efficiency of $\hat{\tilde{Y}}_{r2}$ over $\hat{\tilde{Y}}_{r1}$, given a certain amount of auxiliary information. The performance of the optimal estimator appears to be good, even in finite size samples, and its use profitable, provided that the value of $\lambda(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$ is big enough to compensate for its greater instability. In fact, the empirical results confirm a greater instability in the optimal estimator, especially with asymmetric populations. Further empirical evidence is needed to evaluate its stability when the auxiliary variable is multivariate and to establish when a sample is large enough to overcome the problem.

In order to use the information provided by $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$ within the same survey, the distributional properties of this sample statistic and of the sample dependent regression

estimator, which seems to perform well in the empirical study, have to be studied in more detail. In particular, the distribution of $\hat{\lambda}(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$ when its true value is zero will be useful for choosing the threshold over which shifting from $\hat{\tilde{Y}}_{r1}$ to $\hat{\tilde{Y}}_{r2}$ is truly profitable. Besides working with larger sample sizes, the instability problem of this statistic can be addressed by looking for more stable, consistent estimators of the variances and covariances appearing in $\lambda(\hat{\tilde{Y}}_{r1}, \hat{\tilde{Y}}_{r2})$. Furthermore, since in most practical situations there is more than one variable of interest, in order to apply the same weights to all variable, the optimal estimator should be chosen on the grounds of an averaged λ -measure across the main survey variables, and such an average is more stable than single λ -measures.

ACKNOWLEDGEMENTS

This research was partially funded by a grant of M.U.R.S.T. of Italy. The author is grateful to an associate editor and the referees for their constructive comments which greatly improved upon an earlier version of this paper.

REFERENCES

- CASADY, R.J., and VALLIANT, R. (1993). Conditional properties of post-stratified estimators under normal theory. *Survey Methodology*, 19, 183-192.
- COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons.
- DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ELTINGE, J.L., and JANG, D.S. (1996). Stability measures for variance component estimators under a stratified multistage design. *Survey Methodology*, 22, 157-165.
- ESTEVAO, V., HIDIROGLOU, M.A., and SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- ISAKI, C.T., and FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- LEHTONEN, R., and PAHKINEN, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley.
- MONTANARI, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *International Statistical Review*, 55, 191-202.
- RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of Official Statistics*, 10, 153-165.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

Combining Multiple Frames to Estimate Population Size and Totals

DAWN E. HAINES and KENNETH H. POLLOCK¹

ABSTRACT

Efficient estimates of population size and totals based on information from multiple list frames and an independent area frame are considered. This work is an extension of the methodology proposed by Hartley (1962) which considers two general frames. A main disadvantage of list frames is that they are typically incomplete. In this paper, we propose several methods to address frame deficiencies. A joint list-area sampling design incorporates multiple frames and achieves full coverage of the target population. For each combination of frames, we present the appropriate notation, likelihood function, and parameter estimators. Results from a simulation study that compares the various properties of the proposed estimators are also presented.

KEY WORDS: Incomplete frame; Capture-recapture sampling; Screening estimator; Dual frame methodology; Multiple frame estimation.

1. INTRODUCTION

In classical sampling theory, it is assumed that a complete frame exists. In practice, however, this assumption is often violated. Frame imperfections such as omissions, duplications, and inaccurate recordings are almost inevitable in any large data collection operation (Hansen, Hurwitz and Madow 1953). Information collected from list and area frames is used to obtain estimates of the unknown population size and totals. For example, an ecologist or wildlife biologist may use one list and one area frame sample to estimate the number of bald eagle nests in a given region. The U.S. Bureau of the Census uses dual system estimation to measure decennial census undercounts. Darroch, Fienberg, Glonek and Junker (1993) describe a three-sample multiple-capture approach to estimating population size when inclusion probabilities are heterogeneous. In addition, state agriculture officials may be interested in estimating the number of hog farms and the total number of hogs in North Carolina. Typically, information from multiple information sources is combined to estimate population sizes and totals.

List frames are physical listings of sampling units in the target population. These are constructed over the years using information from scientists as well as city, county, state, and federal agencies. Items found on a list frame can include, but are not limited to, names, addresses, telephone numbers, social security numbers, or physical descriptions of location. These and other miscellaneous stratification variables are used to identify persons, animals, businesses, or other establishments. When estimating the number of bald eagle nests in a region, we construct this year's list frame using information from last year's list frame. With

the addition of new eagle nests, last year's list frame becomes quickly outdated and incomplete. Because of this incompleteness, estimates based solely on list frames typically underestimate the true population size. Supplementing available information with an area frame sample may provide an efficient estimation of the population size and totals.

An area frame is a collection of geographical areas defined by identifiable boundaries. The entire area in which data are collected is divided into mutually exclusive and exhaustive sampling units called segments. The segments are usually stratified according to a characteristic of interest. Once a stratified random sample of segments is drawn, enumerators visit the sampled segments and record measurements on all reporting units contained therein.

The National Agricultural Statistics Service (NASS) currently employs a multi-frame approach for its sampling and estimation of numerous agricultural commodities. Fecso, Tortora and Vogel (1986) provide a review of sampling frames for the agricultural sector of the United States while Nealon (1984) details the multiple and area frame estimators used by the U.S. Department of Agriculture. Kott and Vogel (1995) provide a general overview of multiple frame surveys.

In Section 2, we consider estimation based on information from two or more independent list frames. We show how these methods are related to capture-recapture methods. In Section 3, we consider more efficient estimators of population size and totals when information from an independent area frame sample is available. We extend these methods to the case of dependent list frames in Section 4. Results from a simulation study that compare different estimators are summarized in Section 5. Finally,

¹ Dawn E. Haines, U.S. Bureau of the Census, Washington, DC 20233; Kenneth H. Pollock, North Carolina State University, Department of Statistics, Box 8203, Raleigh, NC 27695-8203, U.S.A.

Section 6 summarizes our results and discusses future directions for research.

2. MULTIPLE LIST FRAMES

2.1 Population Size Estimation

List frames used to estimate population size are usually incomplete and do not cover the entire population. One solution to the incomplete list frame problem is to merge two or more incomplete list frames. Combining multiple list frames may result in improved coverage of the target population, and thus, may provide better estimators. In the case of multiple list frames, it is commonly assumed that each element in the population has the same probability of being included on a given list frame. Hence, the list frame elements themselves constitute our “samples.” For example, individuals may decide independently whether or not to list their telephone numbers in the telephone directory with equal probability. In the case of bald eagle nests, this year’s list frame is constructed based on last year’s nest sightings. If we assume that the probability of a nest being sighted is the same for all nests, then the above assumption is valid. Finally, the assumption is also valid in capture-recapture experiments where the first list frame consists of all animals captured on the first sampling occasion and the second list frame consists of all animals captured on the second sampling occasion. This scenario corresponds to Model M_l in the capture-recapture literature. See Otis, Burnham, White and Anderson (1978) for details. Model M_l assumes all animals in the population are equally at risk to capture on each sampling occasion, but this probability can vary over different sampling occasions.

To begin, we consider the case of two independent list frames, B_1 and B_2 . Suppose B_1 has size N_{B_1} and B_2 has size N_{B_2} . Let domain $b_1(b_2)$ consist of those $N_{b_1}(N_{b_2})$ elements that belong only to frame $B_1(B_2)$ and domain b_1b_2 contain $N_{b_1b_2}$ units that belong to both frames. The final domain includes existing target population elements that are not included on either list frame. Its size is $N - N_{b_1} - N_{b_2} - N_{b_1b_2}$. Domain notation for list frames B_1 and B_2 is presented in Table 1. Note that every element in every frame must be categorized into a domain without error. Errors in domain determination are serious and cannot be corrected at a later time. These errors are not considered in the estimation phase and thus are regarded as nonsampling errors. Nealon (1984) claims that domain determination is the single largest source of nonsampling error in multiple frame designs (Kott and Vogel 1995).

Let the probability that a population element is included on frame $B_1(B_2)$ be $p_{B_1}(p_{B_2})$. Since list frames B_1 and B_2 are assumed to be independent, the probability of an element belonging to domain b_1 is $p_{b_1} = p_{B_1}(1 - p_{B_2})$. The remaining domain probabilities are defined similarly. The population size N and the inclusion probabilities p_{B_1} and

p_{B_2} are unknown parameters. The likelihood function is given by

$$\mathcal{L}(p_{B_1}, p_{B_2}, N | N_{b_1}, N_{b_2}, N_{b_1b_2}) = \binom{N}{N_{b_1}, N_{b_2}, N_{b_1b_2}} * p_{B_1}^{N_{b_1}} p_{B_2}^{N_{b_2}} (1 - p_{B_1})^{N - N_{b_1}} (1 - p_{B_2})^{N - N_{b_2}}. \quad (1)$$

Table 1
Domain Notation for List Frames B_1 and B_2

Domain Size	Domain Probability
N_{b_1}	$p_{b_1} = p_{B_1}(1 - p_{B_2})$
N_{b_2}	$p_{b_2} = (1 - p_{B_1})p_{B_2}$
$N_{b_1b_2}$	$p_{b_1b_2} = p_{B_1}p_{B_2}$
$N - N_{b_1} - N_{b_2} - N_{b_1b_2}$	$1 - p_{b_1} - p_{b_2} - p_{b_1b_2} = (1 - p_{B_1})(1 - p_{B_2})$

Maximum likelihood estimators (MLEs) of the frame inclusion probabilities are obtained by maximizing the logarithm of the likelihood (1). This procedure yields

$$\hat{p}_{B_1} = \frac{N_{B_1}}{\hat{N}} \quad \text{and} \quad \hat{p}_{B_2} = \frac{N_{B_2}}{\hat{N}}, \quad (2)$$

where the MLE \hat{N} is substituted for N . Rather than differentiating the log-likelihood function to approximate the value of N , we employ the “ratio method” of maximizing the likelihood which equates $\mathcal{L}(N)$ to $\mathcal{L}(N - 1)$ (Darroch 1958). This process accounts for the discrete parameter N and yields the equation

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N - 1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - N_{b_2} - N_{b_1b_2})} * (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) = 1. \quad (3)$$

Here we assume that N is large so that

$$\frac{N_{B_1}}{N - 1} \approx \frac{N_{B_1}}{N} \quad \text{and} \quad \frac{N_{B_2}}{N - 1} \approx \frac{N_{B_2}}{N}.$$

Substituting the estimators in (2) into (3) yields

$$\hat{N}_1 = \hat{N} = \frac{N_{B_1}N_{B_2}}{N_{b_1b_2}}. \quad (4)$$

Sekar and Deming (1949) derive an estimate of the variance of (4), given by

$$\hat{V}(\hat{N}_1) = \frac{N_{B_1}N_{B_2}N_{b_1}N_{b_2}}{(N_{b_1b_2})^3}.$$

Substituting (4) into (2) yields the MLEs of p_{B_1} and p_{B_2} ,

$$\hat{p}_{B_1} = \frac{N_{b_1 b_2}}{N_{B_2}} \quad \text{and} \quad \hat{p}_{B_2} = \frac{N_{b_1 b_2}}{N_{B_1}}.$$

The estimator \hat{N}_1 of N in (4) is called the Lincoln-Petersen estimator in closed population capture-recapture models. The elements on list frame B_1 may be considered as the units captured in the first sampling occasion and the elements on list frame B_2 may be viewed as the units captured in the second sampling occasion. The elements in domain $b_1 b_2$ correspond to recaptured elements. With this correspondence, it is easy to see that the likelihood for the population size and capture probabilities for two occasions will be the same as that given in (1). Hence, the MLEs derived for two independent list frames will be the same as the corresponding MLEs for the capture-recapture model with two sampling occasions.

Extending these ideas, we contend that combining k independent list frames is directly related to having k sampling occasions under Model M_t in closed population capture-recapture models, where $t = k$ (Otis *et al.* 1978). The general likelihood function for k independent list frames, B_1, B_2, \dots, B_k , has the form

$$\mathcal{L}(p_{B_1}, \dots, p_{B_k}, N | N_{b_1}, \dots, N_{b_1 \dots b_k}) = \binom{N}{N_{b_1}, \dots, N_{b_1 \dots b_k}} \prod_{l=1}^k p_{B_l}^{N_{B_l}} (1 - p_{B_l})^{N - N_{B_l}}, \quad (5)$$

which has exactly the same structure as the likelihood introduced by Darroch (1958) and is discussed in great detail by Otis *et al.* (1978) and Seber (1982). The form of the estimated frame inclusion probabilities is

$$\hat{p}_{B_l} = \frac{N_{B_l}}{\hat{N}}, \quad l = 1, \dots, k. \quad (6)$$

Values of \hat{N} are obtained by numerically solving the $(k - 1)$ degree polynomial in \hat{N} resulting from the equality

$$\frac{\mathcal{L}(N)}{\mathcal{L}(N - 1)} = \frac{\hat{N}}{(\hat{N} - N_{b_1} - \dots - N_{b_1 \dots b_k})} * (1 - \hat{p}_{B_1}) \dots (1 - \hat{p}_{B_k}) = 1. \quad (7)$$

We then select as \hat{N} as the root that maximizes the value of the likelihood function (5). Substituting this root into (6) yields MLEs of the k frame inclusion probabilities.

2.2 Population Total Estimation

Suppose the measured y_i values are available for all units on the k independent list frames. The estimated probability that the first element is included on at least one of the k list frames is

$$\hat{\pi}_1 = \hat{P}\left[\cup_{l=1}^k B_l\right] = 1 - (1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \dots (1 - \hat{p}_{B_k}),$$

where $\hat{p}_{B_l} = N_{B_l}/\hat{N}$ and \hat{N} is the MLE of N obtained from (7). From equation (7),

$$\frac{\hat{N}}{(\hat{N} - N_{b_1} - \dots - N_{b_1 \dots b_k})} (1 - \hat{\pi}_1) = 1$$

which simplifies to

$$\hat{\pi}_1 = \frac{N_{b_1} + \dots + N_{b_1 \dots b_k}}{\hat{N}}.$$

An estimated Horvitz and Thompson (1952) estimator of the population total is

$$\begin{aligned} \hat{Y}_{H-T} &= \frac{1}{\hat{\pi}_1} \sum_{i \in B_1 \cup \dots \cup B_k} y_i \\ &= \frac{\hat{N}}{N_{b_1} + \dots + N_{b_1 \dots b_k}} \sum_{i \in B_1 \cup \dots \cup B_k} y_i = \hat{N} \bar{Y}_L, \end{aligned}$$

where \bar{Y}_L is the mean of distinct elements on the list frames. Thus, for k independent list frames, the estimated Horvitz-Thompson estimator coincides with the population total estimator proposed by Pollock, Turner and Brown (1994).

In some situations, values of the variable of interest, y_i , are not available for all units on the list frames. If the list frames are large in size, random samples are selected from each list frame and data are collected on those subsampled elements. If there are k list frames, it is possible to define 2^k domains. We consider an extension of Lund's (1968) estimator for the total of all units on the list frames,

$$\hat{Y}_{L,L} = \sum_{l=1}^{2^k-1} N_l \bar{y}_l,$$

which is a weighted sum of $2^k - 1$ domain means, \bar{y}_l . The weights are given by the domain sizes. Further, the population total estimator is

$$\hat{Y} = \hat{N} \frac{\hat{Y}_{L,L}}{\sum_{l=1}^{2^k-1} N_l}.$$

3. MULTIPLE LISTS PLUS AN AREA FRAME

3.1 Population Size Estimation

Joining multiple, individual list frames with an area frame sample is a solution to overcoming list frame deficiencies. Assume that the geographical area of interest is

subdivided into U_A segments. Also, assume that a simple random sample of u_A segments is selected from U_A segments that cover the entire population. Therefore, the probability of a segment being selected is $p_A = u_A/U_A$. In some surveys, it is possible to subdivide the region into approximately equally-sized segments. In such cases the segment selection probability corresponds approximately to the proportion of area sampled. The inclusion of an area frame provides completeness of the target population (Hartley 1962). We assume that each reporting unit belongs to exactly one segment. Once a segment is selected, all reporting units within the segment are observed. For example, when estimating the number of bald eagle nests, each nest belongs to one and only one segment. However, this assumption is not always valid. Consider the case where a hog farm crosses segment boundaries. In this case, population elements may be associated with more than one segment. To address this problem, association rules linking population elements to segments are established at the estimation stage. See Faulkenberry and Garoui (1991) for more detail. The National Agricultural Statistics Service implements three correspondence rules that map elements in the population to sampled segments. The open, closed, and weighted segment estimators are described in Nealon (1984). Another related reference is Sirken (1970).

Consider the case of k independent list frames plus an area frame. The population size, N , and the list frame inclusion probabilities, p_{B_i} , $i = 1, \dots, k$, are unknown parameters. The area frame inclusion probability $p_A = u_A/U_A$ is known. The likelihood function has the form

$$\begin{aligned} \mathcal{L}(p_{B_1}, \dots, p_{B_k}, N | p_A, n_a, n_{ab_1}, \dots, n_{ab_1 \dots b_k}, N_{b_1}, \dots, N_{b_1 \dots b_k}) \\ = \binom{N}{n_a, n_{ab_1}, \dots, n_{ab_1 \dots b_k}, N_{b_1}, \dots, N_{b_1 \dots b_k}} p_A^{n_a} (1 - p_A)^{N - n_a} \\ \prod_{l=1}^k p_{B_l}^{N_{B_l}} (1 - p_{B_l})^{N - N_{B_l}}, \end{aligned}$$

where n_A is the total number of elements in the u_A sampled area segments and n_a is the number of elements in the u_A sampled area segments which do not belong to any list frames. Similarly, $n_{ab_1}, \dots, n_{ab_1 \dots b_k}, N_{b_1}, \dots, N_{b_1 \dots b_k}$ are defined as the sizes of different domains. It is important to emphasize that the inclusion of an area frame may cause the value of N_{b_1} to change. N_{b_1} now corresponds to the number of elements on list frame B_1 which are not in the u_A selected area segments and not on any other list frame.

The MLEs of the parameters are given by $\hat{p}_{B_i} = N_{B_i}/\hat{N}$, where \hat{N} is a solution to the k -th degree polynomial

$$\begin{aligned} \hat{N}(1 - p_A)(1 - \hat{p}_{B_1}) \dots (1 - \hat{p}_{B_k}) = \\ (\hat{N} - n_a - n_{ab_1} - \dots - n_{ab_1 \dots b_k} - N_{b_1} - \dots - N_{b_1 \dots b_k}). \end{aligned} \quad (8)$$

Numerical methods are essential for solving (8) for the MLE \hat{N} of N . Among the k roots of (8), we select \hat{N} that maximizes the likelihood.

Applying this methodology to one list frame and one area frame, we obtain

$$\hat{N} = N_{B_1} + \frac{n_a}{p_A}. \quad (9)$$

This estimator is also known as the screening estimator (Kott and Vogel 1995). The screening estimator categorizes elements into two distinct groups. The first group contains elements which belong to both the list and area frames and is called the overlap domain. Since it is assumed that all elements on a list frame belong to the area frame, the size of the overlap domain coincides with the number of elements on frame B_1 and has the value N_{B_1} . The second group contains elements in the area frame not included on the list frame(s) and is referred to as the nonoverlap domain. The size of the nonoverlap domain is an unobserved random quantity, N_a . The term n_a is the number of elements found in the u_A area segments which are not included on the list frame(s) following a specific association rule. An estimated value of N_a is n_a/p_A . Hence, an estimate of the population size is given by \hat{N} in (9). The resulting MLE of p_{B_1} is

$$\hat{p}_{B_1} = \frac{N_{B_1}}{N_{B_1} + \frac{n_a}{p_A}}.$$

When multiple list frames are available, it is possible to combine them into a single list frame and use the above estimator to obtain an estimate of N . That is, consider the screening estimator

$$\begin{aligned} \hat{N}_2 = \hat{N} = N_{B_1 \cup \dots \cup B_k} + \frac{n_a}{p_A} = N_{b_1} + \dots + N_{b_k} + \\ N_{b_1 b_2} + \dots + N_{b_1 \dots b_k} + \frac{n_a}{p_A}. \end{aligned} \quad (10)$$

Note that the screening estimator \hat{N}_2 is appropriate even when the list frames are *not* independent of each other. We discuss this further in Section 4.

Using this methodology for one area and two independent list frames yields the likelihood

$$\begin{aligned} \mathcal{L}(p_{B_1}, p_{B_2}, N | p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}) = \\ \binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2}} p_A^{n_a} p_{B_1}^{N_{B_1}} p_{B_2}^{N_{B_2}} \\ (1 - p_A)^{N - n_a} (1 - p_{B_1})^{N - N_{B_1}} (1 - p_{B_2})^{N - N_{B_2}}. \end{aligned}$$

The MLE of N is

$$\hat{N}_3 = \hat{N} = (2p_A)^{-1} * \frac{[(N_{B_1} + N_{B_2})p_A + (n_a - N_{b_1b_2} - n_{ab_1b_2})] + (2p_A)^{-1} \sqrt{[(N_{B_1} + N_{B_2})p_A + (n_a - N_{b_1b_2} - n_{ab_1b_2})]^2 + 4p_A(1-p_A)N_{B_1}N_{B_2}}}{(11)}$$

where $n_{ab_1b_2}$ denotes the number of elements included in the u_A sampled area segments that belong to both list frames. An estimate of the variance of \hat{N}_3 may be obtained using the Taylor series approximation of (11) and the asymptotic distribution of $(N_{B_1}, N_{B_2}, n_a, N_{b_1b_2}, n_{ab_1b_2})$.

3.2 Population Total Estimation

When y_i 's are available for all elements on k independent list frames and for a sample of segments from an area frame, we consider an estimated Horvitz-Thompson estimator to estimate the population total. Recall that we assume the following:

1. The probability that a unit is included on the i -th list frame, p_{B_i} , is the same for all units.
2. The event that a unit is included on one frame is independent of its inclusion on another frame.
3. The probability that a unit is included in the area frame sample of u_A segments is $p_A = u_A / U_A$.

Since we consider the case where population units belong to exactly one area segment and all units within a sampled segment are observed, the third assumption is valid. Hence, the probability the i -th element is on at least one of the k list frames and/or the area frame sample is

$$\hat{\pi}_1 = 1 - (1 - p_A)(1 - \hat{p}_{B_1})(1 - \hat{p}_{B_2}) \cdots (1 - \hat{p}_{B_k}) = \frac{n_a + n_{ab_1} + \cdots + N_{b_1 \dots b_k}}{\hat{N}}.$$

The estimated Horvitz-Thompson population total estimator is

$$\hat{Y}_{H-T} = \frac{\hat{N}}{n_a + n_{ab_1} + \cdots + N_{b_1 \dots b_k}} \sum_{i \in \text{sample}} y_i = \hat{N} \bar{y}_L,$$

where \bar{y}_L is the mean of the distinct elements on list frames B_1, \dots, B_k and the elements in the area frame sample.

We can also use the screening estimator to estimate the population total. The known overlap domain total is combined with an estimator of the nonoverlap domain (NOL) total to yield $\hat{Y}_S = Y_L + \sum_{i \in \text{NOL}} y_i / p_A$. The NOL domain consists of elements on the area frame that are not on any of the list frames and $Y_L = Y_{B_1 \cup \dots \cup B_k}$ is the total of the

distinct units on the k list frames. In the subsampling case, we may replace Y_L in \hat{Y}_S by Lund's estimator, given by

$$\hat{Y}_{L,L} = N_{b_1} \bar{y}_{b_1} + \cdots + N_{b_k} \bar{y}_{b_k} + N_{b_1b_2} \bar{y}_{b_1b_2} + \cdots + N_{b_1 \dots b_k} \bar{y}_{b_1 \dots b_k}.$$

4. DEPENDENT LIST FRAMES

We now consider the case where dependencies exist among list frames but where area and list frames remain independent. In capture-recapture experiments, for example, the probability an animal is captured on the second sampling occasion may depend on whether it was captured on the first sampling occasion. See Fienberg (1972), Cormack (1989), Wolter (1990), Pollock, Hines, and Nichols (1984), Huggins (1989), and Alho (1990) for specific examples.

We consider the case where we have two list frames, B_1 and B_2 , that are dependent. Let p_{11} denote the probability of being included on both list frames. If B_1 and B_2 are independent, then $p_{11} = p_{B_1} p_{B_2}$ where p_{B_1} and p_{B_2} are inclusion probabilities for B_1 and B_2 , respectively. Define $p_{10}(p_{01})$ as the probability of being included on frame $B_1(B_2)$ but not on frame $B_2(B_1)$. The probability of exclusion from both list frames is denoted by $p_{00} = 1 - p_{B_1} - p_{B_2} + p_{11}$.

The likelihood function is given by

$$\begin{aligned} \mathcal{L}(p_{B_1}, p_{B_2}, p_{11}, N | p_A, n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}) \\ = \binom{N}{n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1b_2}, n_{ab_1b_2}} p_A^{n_a} (1 - p_A)^{N - n_a} \\ (p_{B_1} - p_{11})^{N_{b_1} + n_{ab_1}} (p_{B_2} - p_{11})^{N_{b_2} + n_{ab_2}} p_{11}^{N_{b_1b_2} + n_{ab_1b_2}} \\ (1 - p_{B_1} - p_{B_2} + p_{11})^{N - N_{b_1} - N_{b_2} - n_{ab_1} - n_{ab_2} - N_{b_1b_2} - n_{ab_1b_2}}. \end{aligned} \quad (12)$$

Maximizing (12) with respect to p_{B_1}, p_{B_2}, p_{11} and N leads to the approximate solution

$$\hat{N} = N_{b_1} + N_{b_2} + n_{ab_1} + n_{ab_2} + N_{b_1b_2} + n_{ab_1b_2} + \frac{n_a}{p_A},$$

which coincides with the screening estimator \hat{N}_2 . That is, \hat{N} is also the estimator that is obtained by pooling the two list frames into a single list frame where the duplications are eliminated and the nonoverlap domain size is estimated using the area frame sample. Also, it can be shown that the two-stage maximum likelihood procedure of Sanathanan (1972) leads to:

$$\hat{N} = \frac{n_a + N_{B_1 \cup B_2}}{p_A + (1 - p_A) \frac{N_{B_1 \cup B_2}}{\hat{N}_2}} = \hat{N}_2.$$

Thus, the maximum likelihood estimator and Sanathanan's estimator both coincide with the screening estimator. If information from two dependent list frames is available and the nature of the dependency is unknown, then we cannot estimate the individual parameters. When information from an independent area frame is available, all parameters are estimable. However, for estimating N , $N_{B_1 \cup B_2}$ is sufficient and no additional information is gained from \hat{N}_{B_1} , \hat{N}_{B_2} , and $\hat{N}_{b_1 b_2}$.

Methods are available for modeling the dependence among k list frames when estimating population size and totals. Additional population information or information from an independent area frame is needed to accurately model the dependence. Fienberg (1972) and Cormack (1989) consider constrained log-linear models to model the dependence. On the other hand, Wolter (1990) uses external constraints such as a known sex ratio to estimate the population size in the dependence case. Another technique used is to model the inclusion probabilities as a function of the covariates. Alho, Mulry, Wurdeman and Kim (1993) use a conditional logistic regression model to estimate the probability of being enumerated in a census and apply the model to the 1990 Post-Enumeration Survey. The role of auxiliary variables in capture-recapture experiments with unequal capture probabilities is addressed in Pollock *et al.* (1984), Huggins (1989), and Alho (1990).

5. SIMULATION STUDY

We conduct a simulation study to assess the overall efficiency of different population size estimators for the special case of two list frames plus an area frame. This is the most feasible combination of sampling frames for real survey problems.

5.1 Design of the Study

In order to study both dependent and independent cases, we define the parameter θ that reflects the dependence structure between list frames B_1 and B_2 . It has the same form as the odds ratio and is written formally as

$$\theta = \frac{P_{00}P_{11}}{P_{01}P_{10}}.$$

In the case of two list frames, the value of θ determines a unique solution for p_{11} . Our study varies the following factors:

Factor	Levels	Definition
N	500, 5000	Population size
p_A	0.05, 0.10, 0.20	Inclusion probability for area frame A
$p_{B_1} (= p_{B_2})$	0.7, 0.9	Inclusion probability for list frame $B_1 (B_2)$
θ	0.5, 1.0, 1.5, 2.0	Odds ratio

For each parametric combination, we generate data $(n_a, N_{b_1}, N_{b_2}, n_{ab_1}, n_{ab_2}, N_{b_1 b_2}, n_{ab_1 b_2})$. One thousand Monte Carlo replications are generated for each parametric combination.

5.2 Estimators

We compare four population size estimators, \hat{N}_1 , \hat{N}_2 , \hat{N}_3 , and \hat{N}_4 . \hat{N}_1 is the Lincoln-Petersen estimator which does not incorporate area frame information. The estimator \hat{N}_1 is suitable when the list frames are independent. Since the estimator ignores information from the area frame sample, it is expected to be inefficient when information from an area frame is available. The screening estimator, \hat{N}_2 , sums the overlap and nonoverlap domain estimates and is particularly suitable for the dependent list frame case. The third estimator, \hat{N}_3 , is derived from the full, independent sampling frame likelihood function. This estimator exploits the information contained in the area and list frames and the fact that the list frames are independent ($\theta = 1$).

We expect \hat{N}_3 to be the best estimator when list frames B_1 and B_2 are independent whereas we expect \hat{N}_2 to be the best estimator in the dependent case. As a result, we also consider a pre-test estimator that tests for independence of the list frames. We define \hat{N}_4 to be \hat{N}_2 if there is strong evidence to believe that frames B_1 and B_2 are not independent. Otherwise, we take $\hat{N}_4 = \hat{N}_3$. Formally,

$$\hat{N}_4 = \begin{cases} \hat{N}_2 & \text{if GOF} > \chi_{1, 0.05}^2 = 3.84 \\ \hat{N}_3 & \text{otherwise,} \end{cases}$$

where GOF is the chi-square goodness-of-fit test statistic for testing $H_0: \theta = 1$ and is derived from the following two-way table.

	In B_1	Not In B_1	
In B_2	$n_{ab_1 b_2}$	n_{ab_2}	$n_{A \cap B_2}$
Not In B_2	n_{ab_1}	n_a	$n_{A \cap B_2'}$
	$n_{A \cap B_1}$	$n_{A \cap B_1'}$	n_A

Figure 1. Classification of Sampled Area Frame Elements

Figure 1 categorizes the n_A elements according to their presence on or absence from list frames B_1 and B_2 .

5.3 Comparing the Estimators

Tables 2 and 3 display the percent relative bias and the percent relative root mean square error of the estimates $\hat{N}_1, \hat{N}_2, \hat{N}_3$, and \hat{N}_4 for population sizes of 500 and 5000, respectively. We scale the bias and the root mean square error by N in order to directly compare estimators based on different population sizes. A comparison of \hat{N}_1 with \hat{N}_3 shows the benefit of drawing an area frame sample. In practice, these benefits depend on the relative cost of the area frame sample. In this study, we do not take sampling costs into account. The probability of being included on both list frames, p_{11} , is given in parentheses in the θ column. When $p_B = p_C = .9$, p_{11} must lie between .8 and .9. However, for θ ranging from .5 to 2, p_{11} varied only from .806 to .817.

The estimator \hat{N}_2 is unbiased for N and has the smallest percent relative bias. The estimators \hat{N}_1 and \hat{N}_3 are asymptotically consistent for N and yield biases close to 0 when $\theta = 1$. On the other hand, \hat{N}_1 and \hat{N}_3 have large biases when $\theta \neq 1$. The percent relative bias of \hat{N}_4 is smaller than that of \hat{N}_3 but it is not close to zero. The bias does not change significantly as p_A increases from .05 to .10 to .20.

When $N = 500$ and $p_B = p_C = .9$, \hat{N}_3 has the smallest percent relative root mean square error (% RRMSE). This is partly due to the fact that the limited range of p_{11} values is similar to the p_{11} value for the independence case (.810). The % RRMSE for \hat{N}_3 is 40 - 50 % smaller than that of \hat{N}_2 . On the other hand, the % RRMSE of \hat{N}_3 is only 15 - 30 % smaller than that of \hat{N}_1 . Therefore, when the list frames have very high inclusion probabilities, both \hat{N}_1 and \hat{N}_3 are much better than \hat{N}_2 . Additionally, if area frame sampling costs are high, \hat{N}_1 may be a reasonable alternative estimator to \hat{N}_3 . When $N = 500$ and $p_B = p_C = .7$, \hat{N}_3 has the smallest % RRMSE for the independence case. When $\theta = 2$, \hat{N}_2 has the smallest % RRMSE. If $N = 5000$ and $p_B = .7$, \hat{N}_3 has the smallest % RRMSE for only $\theta = 1$. For all other θ values, \hat{N}_2 yields the smallest % RRMSE. In all cases, \hat{N}_3 has very small variance and most of the % RRMSE is due to the bias in \hat{N}_3 . For $\theta < 1$, \hat{N}_3 tends to have positive bias while for $\theta > 1$, \hat{N}_3 has negative bias. For the case of $N = 5000$ and $p_B = .9$, \hat{N}_3 has the smallest % RRMSE for $\theta = 1$. \hat{N}_2 has the smallest % RRMSE for $\theta = .5$ and 2. For $\theta = 1.5$, there is no best estimator with respect to % RRMSE.

As expected, the percent relative root mean square errors of \hat{N}_2, \hat{N}_3 , and \hat{N}_4 decrease as the value of p_A increases. Thus, as the area frame information increases, the % RRMSE decreases. Also, as the population size increases from 500 to 5000, the % RRMSE decreases. Since the values of p_A in our simulation are small, \hat{N}_2 has a large variance. On the other hand, even though \hat{N}_3 is biased, it has a very small standard error and results in a smaller % RRMSE. The estimator \hat{N}_4 reduces the bias of \hat{N}_3

but has a large standard error. Hence, \hat{N}_4 is not a particularly beneficial estimator. For larger values of θ and p_A , we expect \hat{N}_2 to perform better than \hat{N}_3 . For the values of θ and p_A we considered, we recommend \hat{N}_3 over other estimators.

The value of % RRMSE for \hat{N}_4 is between that of \hat{N}_2 and \hat{N}_3 in most cases. We write the estimator \hat{N}_4 as $\hat{N}_4 = \delta \hat{N}_2 + (1 - \delta) \hat{N}_3$, where $\delta = 0$ or 1 based on the results of the goodness-of-fit test. The % RRMSE and % RBias of \hat{N}_4 need not be between those of \hat{N}_2 and \hat{N}_3 because δ is not independent of \hat{N}_2 and \hat{N}_3 .

5.4 Limitations of the Study

The goal of our study is to compare the bias, standard error, and mean square error of four population size estimators. We assume that inclusion probabilities for both list frames are identical. Future studies may include unequal inclusion probabilities as well as larger values of θ . Clearly the benefit of \hat{N}_3 over \hat{N}_1 depends on the cost of sampling from an area frame. Our paper considers only small values of p_A . Small p_A values are associated with a high area frame sampling cost. Even in this case, we observe a significant reduction in % RRMSE and % RBias, thereby justifying the use of \hat{N}_3 over \hat{N}_1 . We do not consider an objective function which incorporates sampling costs, % RRMSE, and % RBias.

Throughout this paper, we assume that all units have the same probability of being included on a given list frame. Haines (1997) considers the case where the inclusion probabilities are modeled as a function of a covariate. When inclusion probabilities are heterogeneous, larger units may have a higher list frame inclusion probability than smaller units. Heterogeneous inclusion probabilities play an important role in estimating population totals when the response variable has a highly skewed distribution or has rare values. Haines (1997) also presents two stratification procedures that are useful when area and list frames are stratified on the same variable. These results will be presented in future publications.

6. DISCUSSION

The primary focus of this paper is population size estimation based on several sampling frames. Information from area and/or list frame(s) is collected and combined to obtain various estimators. We derive population size estimators when information is available only on k independent list frames and also when information is available on an area frame sample in addition to the list frames. We conduct a simulation study to compare the performance of the estimators in the special case of two list frames plus an area frame. Based on our simulation study, we recommend the estimator derived from the full, independent likelihood, \hat{N}_3 , for the case where the list

Table 2
Simulation Results for $N = 500$

p_B	θ		p_A					
			.05		.10		.20	
			% RBias	% RRMSE	% RBias	% RRMSE	% RBias	% RRMSE
.7	.5 (.462)	\hat{N}_1	62.30	66.01	60.64	64.04	63.26	66.81
		\hat{N}_2	0.30	49.07	-0.75	32.37	0.85	22.58
		\hat{N}_3	55.52	58.95	48.15	51.15	40.53	43.32
		\hat{N}_4	48.15	58.88	37.88	49.25	24.95	38.80
	1 (.490)	\hat{N}_1	0.47	19.26	1.01	19.08	-0.11	19.45
		\hat{N}_2	0.45	57.34	0.34	39.61	0.88	27.25
		\hat{N}_3	0.43	18.21	0.83	16.93	0.14	15.75
		\hat{N}_4	2.40	27.57	1.39	22.94	0.29	17.96
	1.5 (.508)	\hat{N}_1	-35.60	40.06	-36.48	40.58	-35.69	40.26
		\hat{N}_2	3.11	66.43	-5.08	41.96	0.30	28.79
		\hat{N}_3	-32.07	36.79	-31.01	35.28	-24.04	28.88
		\hat{N}_4	-22.74	47.62	-26.21	37.57	-17.06	30.38
	2 (.522)	\hat{N}_1	-60.07	62.91	-61.31	64.06	-60.41	63.28
		\hat{N}_2	-6.12	66.59	-1.15	46.68	1.67	30.99
		\hat{N}_3	-55.36	58.35	-51.21	54.19	-40.89	43.99
		\hat{N}_4	-41.39	63.79	-34.79	55.45	-18.60	41.35
	.9 (.806)	\hat{N}_1	5.37	6.79	5.27	6.63	5.59	6.97
		\hat{N}_2	0.08	14.78	-0.06	10.17	-0.06	6.55
		\hat{N}_3	5.04	6.44	4.62	5.93	4.24	5.53
		\hat{N}_4	5.94	9.48	5.03	7.05	4.34	5.72
	1 (.810)	\hat{N}_1	0.30	5.01	0.17	5.01	0.25	4.94
		\hat{N}_2	0.78	20.72	0.41	14.06	-0.06	9.03
		\hat{N}_3	0.33	4.83	0.20	4.68	0.17	4.24
		\hat{N}_4	3.23	13.79	1.88	9.35	1.00	5.98
	1.5 (.814)	\hat{N}_1	-4.29	7.07	-4.39	7.32	-4.55	7.37
		\hat{N}_2	-0.65	21.52	0.35	15.88	0.002	10.27
		\hat{N}_3	-4.07	6.78	-3.83	6.73	-3.49	6.15
		\hat{N}_4	-0.43	13.77	-1.18	10.92	-1.43	8.20
	2 (.817)	\hat{N}_1	-8.28	10.27	-8.40	10.36	-8.33	10.32
		\hat{N}_2	-0.29	25.59	0.39	17.66	0.35	11.41
		\hat{N}_3	-7.80	9.82	-7.35	9.38	-6.30	8.20
		\hat{N}_4	-2.52	17.96	-3.10	14.02	-2.73	10.33

Table 3
Simulation Results for $N = 5000$

P_B	θ		P_A					
			.05		.10		.20	
			% RBias	% RRMSE	% RBias	% RRMSE	% RBias	% RRMSE
.7	.5 (.462)	\hat{N}_1	61.47	61.82	61.39	61.76	61.69	62.04
		\hat{N}_2	-0.18	15.78	0.26	10.65	-0.15	6.72
		\hat{N}_3	54.84	55.17	49.06	49.38	39.38	39.65
		\hat{N}_4	19.73	38.12	4.77	19.52	-0.01	7.21
	1 (.490)	\hat{N}_1	-0.28	6.14	-0.13	5.99	0.35	6.15
		\hat{N}_2	0.43	18.14	0.47	12.85	-0.20	8.34
		\hat{N}_3	-0.22	5.82	-0.03	5.35	0.16	4.88
		\hat{N}_4	0.26	9.82	-0.04	7.44	0.11	5.95
	1.5 (.508)	\hat{N}_1	-36.21	36.68	-36.29	36.78	-35.90	36.38
		\hat{N}_2	0.41	20.39	-0.16	14.21	0.39	9.55
		\hat{N}_3	-32.87	33.37	-29.97	30.49	-24.13	24.66
		\hat{N}_4	-19.11	31.15	-11.51	23.92	-3.12	14.03
	2 (.522)	\hat{N}_1	-61.04	61.3	-60.53	60.81	-60.64	60.92
		\hat{N}_2	0.40	20.09	0.60	15.43	0.31	9.67
		\hat{N}_3	-55.69	55.96	-50.24	50.55	-41.46	41.76
		\hat{N}_4	-14.10	36.31	-2.34	20.96	0.26	9.84
.9	0.5 (.806)	\hat{N}_1	5.56	5.70	5.52	5.67	5.54	5.68
		\hat{N}_2	-0.12	4.55	0.11	3.19	-0.03	2.08
		\hat{N}_3	5.21	5.35	4.86	5.01	4.22	4.35
		\hat{N}_4	4.97	5.41	3.64	4.88	2.26	3.79
	1 (.810)	\hat{N}_1	-0.02	1.58	0.08	1.55	0.01	1.57
		\hat{N}_2	-0.09	6.16	-0.17	4.08	-0.14	2.79
		\hat{N}_3	-0.03	1.53	0.05	1.48	-0.02	1.35
		\hat{N}_4	0.37	3.19	0.11	2.18	0.09	1.89
	1.5 (.814)	\hat{N}_1	-4.66	5.00	-4.52	4.85	-4.61	4.90
		\hat{N}_2	-0.25	7.54	0.11	4.95	-0.09	3.14
		\hat{N}_3	-4.39	4.73	-3.96	4.32	-3.55	3.85
		\hat{N}_4	-2.50	6.31	-2.26	5.02	-1.84	3.82
	2 (.817)	\hat{N}_1	-8.45	8.68	-8.38	8.60	-8.46	8.69
		\hat{N}_2	-0.21	7.86	-0.06	5.29	0.01	3.73
		\hat{N}_3	-7.95	8.18	-7.39	7.61	-6.49	6.73
		\hat{N}_4	-3.76	8.80	-2.77	6.99	-1.25	4.97

frames are independent or nearly independent. For the moderate to strong dependence cases, we recommend the screening estimator, \hat{N}_2 .

We also study population total estimation. We consider two scenarios for estimating population totals. In the first case, we assume that observations are available on all units that comprise the list frames. In contrast, the second case assumes that information is available only on subsamples from each of the list frames. We consider an estimated Horvitz-Thompson estimator if list frames are independent and a screening estimator to estimate the population total if the list frames are dependent.

In this paper, our focus is on population size estimation. In practice, one may be interested in estimating population totals for several characteristics based on multi-stage samples involving unequal inclusion probabilities. Relevant papers on this topic include Bankier (1986), Skinner (1991), and Skinner, Holmes, and Holt (1994).

7. ACKNOWLEDGEMENTS

The authors thank the editor and two referees for useful comments on an earlier version of the article. This research was partially funded by the U.S. Geological Survey, Biological Resources Division. Christine Bunck is the BEST Program Manager. The views expressed are attributed to the authors and do not necessarily reflect those of the Census Bureau.

REFERENCES

- ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- ALHO, J.M., MULRY, M.H., WURDEMAN, K., and KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- CORMACK, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.
- DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., and JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.
- FAULKENBERRY, G.D., and GAROUI, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.
- FECISO, R., TORTORA, R.D., and VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.
- FIENBERG, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^k contingency tables. *Biometrika*, 59, 591-603.
- HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Ph.D. thesis, North Carolina State University.
- HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.
- HARTLEY, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- KOTT, P.S., and VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Ed., B.G. Cox). New York: John Wiley & Sons, 185-203.
- LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.
- NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators. Staff Report 80, U.S. Department of Agriculture, Statistical Reporting Service, Washington, DC.
- OTIS, D.L., BURNHAM, K.P., WHITE, G.C., and ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.
- POLLOCK, K.H., HINES, J.E., and NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.
- POLLOCK, K.H., TURNER, S.C., and BROWN, C.A. (1994). Use of capture-recapture techniques to estimate population size and population totals when a complete frame is unavailable. *Survey Methodology*, 20, 117-124.
- SANATHANAN, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43, 1, 142-152.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*, (2nd Edition). New York: Macmillan.
- SEKAR, C.C., and DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- SKINNER, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- SKINNER, C.J., HOLMES, D.J., and HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *International Statistical Review*, 62, 333-347.
- WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

Temporary Mobility and Reporting of Usual Residence

NANCY BATES and ELEANOR R. GERBER¹

ABSTRACT

Temporary mobility is hypothesized to contribute toward within-household coverage error since it may affect an individual's determination of "usual residence" – a concept commonly applied when listing persons as part of a household-based survey or census. This paper explores a typology of temporary mobility patterns and how they relate to the identification of usual residence. Temporary mobility is defined by the pattern of movement away from, but usually back to a single residence over a two-three month reference period. The typology is constructed using two dimensions: the variety of places visited and the frequency of visits made. Using data from the U.S. Living Situation Survey (LSS) conducted in 1993, four types of temporary mobility patterns are identified. In particular, two groups exhibiting patterns of repeat visit behavior were found to contain more of the types of people who tend to be missed during censuses and surveys. Log-linear modeling indicates that temporary mobility patterns are a significant predictor of usual residence, even when controlling for the amount of time spent away and demographic characteristics.

KEY WORDS: Temporary mobility; Usual residence; Household rosters; Coverage.

1. INTRODUCTION

The fundamental challenge in any census of population is the accurate and complete count of every person within that population. Consequently, the extent to which people are missed or undercounted during a census is arguably the most important measure by which it is evaluated. Most censuses and household-based surveys begin with a roster question designed to list all "usual residents" of a household.

Research evaluating the quality of census data suggests that coverage error is a problem. In 1990, the U.S. Post Enumeration Survey (PES) and demographic analyses estimated that the net national undercount was approximately 2% (Hogan 1993; Robinson, Ahmed, Das Gupta and Woodrow 1993). Other research suggests that coverage error in current surveys (such as the U.S. Current Population Survey) is even larger than undercoverage occurring during decennial censuses (Shapiro, Diffendal, Cantor 1993; Chakrabarty 1992; Pennie 1990; Hainer, Hines, Martin and Shapiro 1988). Research by Fein and West (1988) and Shapiro *et al.* (1993) suggest that failure to count all persons within a housing unit is a larger component of total coverage error than failure to count persons as a result of missing a housing unit. Others report that within-household omissions account for about one-third of all census omissions (Ellis 1994; Fay 1989a).

Coverage research also indicates that persons who are undercounted are not randomly distributed among the population. For example, blacks and Hispanics are undercounted at a higher rate than non-Hispanic whites (4.6% and 4.0%, respectively, compared to 0.7%; Hogan 1993). Persons who reside in multi-unit structures (such as apartments) and those who rent are also more likely to be

missed (Griffin and Moriarity 1992; Moriarity and Childers 1993; Ellis 1993).

This paper concentrates on a dimension long hypothesized to contribute to within-household coverage error. This dimension focuses on temporary mobility into and out of a residence over a period of time. Specifically, we examine movement in terms of the number of places a person may visit, the number of visits he/she makes and the amount of time he/she spends there. This analysis examines whether or not mobility may be a factor influencing coverage and indeed be a good indicator of household attachment. We hypothesize that a person's level of mobility tends to influence a household respondent's decision when defining that person as a usual resident and, consequently, someone he/she would or would not include on a census report.

2. BACKGROUND

The movement from one geographical location to another is usually signified by a change of address, movement of possessions and so on. This type of mobility is commonly referred to as geographic mobility. In addition to geographic mobility, there exists a more subtle form of mobility that is not so clearly defined – temporary mobility. Defined here, temporary mobility refers to the temporary and sometimes patterned movement away from a residence and encompasses both long and short, frequent and infrequent overnight stays. This type of mobility has been described as "one of the key features of irregular and complex households" (de la Puente 1993). One example of this is found in Haitian immigrant communities where typical household structure consists of a relatively

¹ Nancy Bates, Office of the Director, U.S. Bureau of the Census, Room 2031, Federal Building 3, Washington, DC 20233, and Eleanor R. Gerber, Center for Survey Methods Research, U.S. Bureau of the Census, Room 3133, Federal Building 4, Washington, DC 20233 U.S.A.

permanent "nuclear core" and a more mobile "fluid periphery." The fluid periphery consists of related and non-related newcomers, staying for short periods of time, and members of the household who visit Haiti on a regular basis and can be away weeks or months at a time (Wingerd 1992).

Temporary mobility is not limited to special communities. Many examples can be found in the wider community, including mobility associated with long term business or vacation travel, attendance at college, custody situations, and persons who maintain a presence in one or more households over a given period of time. This mobility in the fluid periphery, or temporary mobility, differs from geographic mobility because it consists of movements away from, but usually back to, a single residence over time. Members of this fluid periphery present conceptual difficulties for respondents in identifying which members to include in a census or survey. Movement of these persons may not involve a permanent change in address, and thus can blur the concept of who is defined as living or staying at a given address.

Given that there is little literature on temporary mobility, studies on geographic mobility and household structure provide a good starting point for forming our hypotheses about temporary mobility. According to the March 1994 Current Population Survey, young adults between 20-24 are reported to have the highest rates of geographic mobility, with one-third having moved between March 1993 and March 1994. Differences by race are also evident with a higher rate of mobility among blacks and Hispanics (19.6% and 22.4%, respectively) compared to whites (16.0%, see Hansen, 1994). Finally, tenure is also closely correlated with geographic mobility – renters were four times more likely than homeowners to have moved between 1993 and 1994. Obviously, these geographic movers share many of the same characteristics as some undercounted populations.

The kind of mobility with which we are concerned may also be a reflection of socioeconomic status. Temporary mobility, transitory situations, and peripheral connection to households can represent a means of adjusting for a lack of resources (Lipton and Estrada 1993). Hudgins and Holmes (1993) suggests that the undercounting of young black males is a result of their social and economic marginality evidenced in part by a lack of stable residences and relatively permanent mailing addresses. One facet of this may involve temporary movement to extended families or "kin" networks in order to receive family or financial assistance. This phenomenon of extended or kin networking among blacks has also been documented extensively by ethnographic studies (Martin and Martin 1985; Stack 1974; Hainer *et al.* 1988). These living arrangements suggest nontraditional (or at least non-nuclear) household formations which could contribute to coverage error, especially if a person participates in kin networks by moving back and forth among them.

Finally, Montoya (1992) describes a very different household composition that is characteristic of some recent

Hispanic immigrant communities. Like kin-network households, they contain people who come and go, however, the members are "loosely tied, ephemeral, and alienated" and often composed of young migrant men who work and sleep in different shifts and have virtually no social ties with one another. Several other ethnographers have identified similar households in other Hispanic communities across the United States (Velasco 1992; Mahler 1993; Romero 1992.) They found that census coverage in such households was often restricted to those individuals who were actually present when the enumerator arrived.

3. METHODOLOGY

Data for this analysis come from the Living Situation Survey (LSS), a survey specifically designed to gather information about household membership, social attachments, mobility and the assignment of usual residence. The LSS was a voluntary survey conducted by the Research Triangle Institute (RTI) and sponsored by the U.S. Census Bureau between May and September of 1993. The sample was stratified to oversample for high and medium minority areas (*i.e.*, greater than 80% black or Hispanic, between 40% and 80% black or Hispanic) and areas containing renters (*i.e.*, greater than 40% renters). To increase the efficiency of the sample design, RTI used housing unit data previously collected from a multistage probability sample used in the 1992 National Household Survey on Drug Abuse (NHSDA).

The first portion of the LSS interview was conducted in-person with the most knowledgeable household respondent, in most cases, the householder (by U.S. Census Bureau definition, this refers to the person in whose name the house is owned or rented). These householders provided a roster and then answered demographic questions for themselves as well as all other listed persons. Through a series of 13 extensive roster probes, the questionnaire rostered "core" household residents but also included many persons having a less permanent presence. Persons with a more tenuous attachment were brought in by asking probes about who had spent the night there during the reference period, who was considered a household member even if they were staying elsewhere, and who considered the residence their permanent address or a place they received mail or phone messages (see Sweet 1994). (The length of the reference period varied depending upon the date of the interview. Reference periods began on the first day of the month two months prior to the interview month and ended on the day of the interview. Accordingly, interviews conducted toward the end of the month had a longer reference period than interviews conducted near the beginning). In total, 999 households were interviewed nationwide. Using the broad rostering technique, a total of 3,549 people were listed.

The next step in the survey was to weed out rostered individuals determined to be only "casual visitors" to the

household. Individuals were defined as casual visitors if: 1) their usual residence was considered by the householder to be someplace other than the sample housing unit *and* 2) they had stayed at the household for one week or less during the reference period. This screening process identified persons from the broad rostering technique who had only a casual attachment to the household. Of the 3,549 persons rostered, 712 were considered to be casual visitors. (Of the 712 casual visitors, 77% were related to the household respondent, 93% were non-Hispanic, 84% were white and 58% were female). For several reasons, casual visitors were ineligible for the remainder of the questionnaire. First, we assumed that casual visitors do not meet the Census Bureau definition of a usual resident at the interview household and second, excluding this group from the bulk of the questionnaire greatly reduced the time and resources required to carry out the survey.

After follow-up for converting refusals and other non-interviews, the final response rate for the household-level portion of the interview was 79.5%. (Follow-up actions included sending refusal conversion letters, having field supervisors call directly, make repeat visits, and re-assign interviewers. Respondents were contacted an average of 1.9 times; nonrespondents an average of 5.9 times). Considering the population, this was considered to be an acceptable rate of response. Nonetheless, since we suspect that nonresponse is highly related to coverage issues such as mobility, it is likely that this level of nonresponse has some effect upon our estimates. More discussion on this is included in the description of the individual questionnaire below.

The next part of the survey was a self-reported individual-level questionnaire. This part of the survey contained questions about temporary mobility as well as self-reported demographics. Respondents were asked if they had stayed overnight at any other place beside the interview household during the reference period. If so, interviewers used a calendar to record each place and the dates stayed. Interviewers also gathered information about the type of each place stayed, the individual's attachment to each place, and the reason(s) for going there.

Each of the householders answered the individual-level questionnaire for himself/herself. Additionally, all rostered persons who had stayed away for eight or more nights during the reference period answered the individual-level questionnaire. All persons identified as college students and persons with no usual residence were also eligible for an individual interview. Finally, the individual questionnaire was also given to a simple random 10% sample of LSS households. Within these households, individual interviews were attempted with each person on the roster, *with the exception of casual visitors*. This somewhat complex selection criterion resulted in a base of persons representing people with a greater-than-casual association to the interview households, all of whom are included in the analyses reported below ($N = 1,451$).

The individual-level portion of the questionnaire had a response rate of 85.3%. The majority of individual interviews were conducted in-person (96%) and most of the adult interviews (89%) were self-reported while all interviews with children were conducted by a knowledgeable proxy. Because the householders answered basic living situation questions and demographic questions for *all* rostered individuals, we had some means for examining the characteristics of the approximately 15% who were selected for the individual questionnaire but did not respond. We found no significant sex or age differences between nonrespondents and respondents but we found that a disproportionate percentage of nonrespondents were black. We also found that nonrespondents were more likely to have spent more than one week away from the interview household than respondents. These findings shed some light on how representative our individual sample is both demographically and with respect to temporary mobility. Because nonrespondents were reported to be away more than respondents, we suspect the potential 'selectivity' bias may have underestimated our mobility measures.

Household and individual-level weights were applied to adjust for the oversampling, the selection criteria for the individual-level survey and for nonresponse (see Lynch, Witt, Branson and Ardini 1993). All analyses were conducted using Contingency Table Analysis for Complex Sample Designs (CPLX), a computer variance estimation program designed to adjust for the LSS's complex sample design effects (see Fay 1989b; 1985).

3.1 Typology of Temporary Mobility

The typology which we present is empirically based. That is, the particular groupings of visits and destinations was derived analytically and not theoretically. Therefore, the categories we identify do not represent groups of persons with identical characteristics or in identical circumstances. Rather the typology should be regarded as an attempt to represent the complex underlying reality involved in mobile living situations. It is our hypothesis that such mobility has an affect on the strength of the social tie between an individual and a particular household, and that these ties influence the judgment of the household respondent in deciding who is a usual resident of the household. Time away, number of visits and number of destinations are an indirect measure of the strength of such ties.

Our typology of temporary mobility was created using two dimensions of overnight movement outside the interview household. The first dimension taps into the variety of places a person visited over the reference period. This provides some idea of how many places other than the interview household that a person might have attachments to. The second dimension taps the frequency of movements outside the interview household by counting the number of times a person left for a period of one or more nights.

The use of these factors as a measure of the strength of attachment to a household is confirmed by ethnographic descriptions of highly mobile living situations. The pattern of movement represented in our typology reflects many different social processes, such as dispersed attachment to extended kin households (Stack 1974; Dressler, Hoepfner and Pitts 1985), immigration patterns (Wingerd 1992), and adaptation to poverty (Hainer 1987; Valentine and Valentine 1971).

The LSS included several exploratory open-ended questions designed to examine respondents' perception of the reasons for their mobility. The questions asked the reasons for going and reasons for return for particular trips. We had hoped that these questions would provide us with a more direct assessment of the underlying social patterns that cause temporary mobility. Unfortunately the answers to these open ended questions were difficult to code without making unwarranted assumptions, largely as a result of the way in which they were expressed. As a result, we did not incorporate these reasons when formulating the typology.

Each "move" was defined as a stay made outside the interview household for at least one night. For example, if a person left to spend three days at a girlfriend's, then moved from there to a relative's for one night before returning to the interview household that person would be assigned as having two total places with two total visits (one visit apiece). Conversely, if a person left to stay overnight at a friend's then returned to the household and then two weeks later returned to the same friend's home for a second visit, that person would be assigned one place with two total visits (two repeat visits). The first example exemplifies a potential bias in this method, that of counting each unique place visited during one extended trip outside the interview household as an independent move (such as a vacation with multiple destinations). On the other hand, this method also captures the movement of "floaters" by counting each separate place visited during one move away from the household as a separate move.

A single mobility measure using various combinations of the number of places and number of moves was constructed. In all, five categories were created with efforts made to identify different patterns of movement by separating out those making repeat visits to the same places. Our first category depicts persons who stayed all nights of the reference period at the interview household and represents persons with no temporary mobility (the "Non-mobile"). The second category consists of persons who, according to the calendar, reported only one visit to one place (the "1-shots"). The "Boomerangs" reflect persons making repeat visits to one place only. The "No-repeats" are characterized as persons who traveled to more than one place, but never the same place twice. And finally, the "Floaters" stayed overnight at several different places, making repeat visits back to at least one of these places (see table 1).

Table 1
Temporary Mobility Typology

Number of Places Visited	Number of Visits				
	0	1	2	3	4
0	Non-mobile				
1		1-Shots	Boomerangs	Boomerangs	Boomerangs
2			No Repeats	Floaters	Floaters
3				No Repeats	Floaters
4					No Repeats

4. CHARACTERISTICS OF MOBILITY TYPES

Table 2 presents the weighted frequencies for the mobility typology. Slightly more than half of the persons administered the individual questionnaire reported no mobility outside the interview household during the reference period. The largest concentration of persons who were mobile fell into the 1-shot category, that is, they reported making only one move outside the interview household to one place (26%, overall). Eleven percent comprised the Boomerang category reporting a more repetitive pattern of two or more visits to a single place while 7% reported the less patterned, yet highly mobile "No repeat" category. The Floaters comprised the smallest group with 4%.

Table 2
Typology of temporary Mobility by Sex and Hard-To-Enumerate (HTE)* Status (Weighted % and standard errors)

MOBILITY TYPE	Total Weighted Percent (s.e. in paren.)	SEX		HTE STATUS	
		MALE	FEMALE	NON-HTE	HTE
Non-mobile	52% (14.0)	40% (13.7)	67% (13.6)	53% (14.3)	38% (7.8)
1-Shots	26% (10.4)	35% (13.9)	16% (7.0)	27% (10.6)	6% (2.9)
Boomerangs	11% (4.0)	15% (5.7)	6% (2.9)	10% (4.1)	21% (9.1)
No Repeats	7% (2.9)	6% (2.4)	8% (4.3)	7% (3.0)	6% (5.4)
Floaters	4% (1.0)	4% (1.3)	3% (1.3)	3% (0.9)	29% (9.9)
Unweighted N	1,451	653	798	1,375	76
Jackknife chi-square**		$\chi^2 = 2.03, p < .05,$ $d.f. = 4$		χ^2 for distribution excluding non-mobile category = 2.14, $p < .05, d.f. = 4$	

* The hard-to-enumerate group includes black and Hispanic males aged 18-29.

** See Fay 1985 for documentation of Jackknife chi-square test for complex samples.

Tables 2 also illustrates selected demographics for the five mobility categories including gender breakouts which illustrate a higher mobility propensity for males than females. Approximately 60% of the males reported at least one visit outside the interview household, which was significantly higher than females at approximately 33%. This gender difference in temporary mobility is much more pronounced than in geographic mobility where the difference between the male and female move rate is only around 1% (17% of the male population moved between 1993 and 1994 compared to 16% for females, see Hansen 1994). This suggests that temporary mobility is more common than geographic mobility and that the demographic characteristics associated with it are different as well. Military travel could explain the gender differences in temporary mobility, as could travel for business with males having a higher active-duty/population ratio and employment/population ratio compared to females (U.S. Department of Labor 1994).

The right side of Table 2 integrates several demographic characteristics to create a subgroup known to have high rates of undercount in previous censuses. This group is comprised of males between 18 and 29 who are black or Hispanic. This subgroup is sometimes referred to as the "hard-to-enumerate" or HTE population. Only a small percentage of the LSS sample met the HTE criteria, but an examination of this group's mobility reveals very different patterns compared to the non-HTE group.

First, the HTE group appears more mobile to begin with – over 60% indicated spending at least one night someplace other than the interview household compared to less than 50% for non-HTEs. Second, the distribution of mobile categories differs significantly by HTE status. The majority of non-HTEs who are mobile are concentrated in the 1-shot category whereas the HTEs who are mobile are more concentrated in the repeat movement categories (Boomerangs and Floaters with 21% and 29%, respectively).

We also examined the distributions for temporary mobility by race (white, black, Hispanic, and other) and age (0-17, 18-29, 30-49, 50+). Overall, temporary mobility did not vary significantly by either, yet some interesting trends were noticeable. A relatively large concentration of Hispanics were found in the No-Repeat category (19%) and blacks in the Floater group (9%). A higher percentage of blacks were Non-mobile (66%) compared to whites (52%), in spite of the fact that blacks have higher rates of geographic mobility than whites. Finally, young adults between 18 and 29 appeared more mobile than other age groups (close to 70% of this age group spent at least one night away from the interview household) and a disproportionate percentage of this group were Floaters (14%). The lack of statistical significance among some of these trends may be an artifact of sample size. Alternatively, temporary mobility may be sufficiently different from geographic mobility such that it does not share the same characteristics of traditional 'movers'.

Another important variable hypothesized to correlate with the pattern of temporary mobility is the amount of time spent away on visits. The U.S. Census Bureau residence rules vary in the use of time as a criterion for usual residence. For example, persons who work in another city during the week but return home on weekends are to be counted at the place where they "live and sleep" the majority of the time – in this case, at the place they live during the week. However, a child living away at boarding school is to be counted at the parent's residence even though he/she probably spends the majority of time at the school.

Likewise, a person staying at a group quarters on Census Day (e.g., a college dorm or a jail) is counted at that place, regardless of their living situation the rest of the year. Gerber (1994) found that respondents also use time to varying degrees when defining household rosters – in certain situations, she found no clear relationship between being rostered and the amount of time spent at a place. Instead, things like household membership and relationship seemed to factor more heavily in the decision-making process.

Nonetheless, it makes intuitive sense that the amount of time spent away plays some part in the householder's determination of where to count someone. In order to see how our mobility categories varied in term of length of time spent away, the sum of the total number of nights spent away during all visits in the reference period was divided by the total number of nights in the reference period and then expressed as a percentage. Table 3 presents this time measure expressed in terms of being away more or less than half of the reference period.

Table 3
Time Spent Away from the Interview Household during the Reference Period (Weighted % and standard errors)

Away 50% of time or more?	1-Shots	Boomerangs	No Repeats	Floaters	Total
No	94% (4.4)	73% (11.5)	98% (1.4)	63% (10.3)	88% (3.6)
Yes	6% (4.4)	27% (11.5)	2% (1.4)	37% (10.3)	12% (3.6)
Unweighted N	314	186	101	134	735

Jackknife chi-square = 1.71, $p < .05$, $d.f. = 3$

Both the Boomerangs and Floaters were more likely than other groups to spend half or more of the reference period someplace other than the interview household. This supports the notion that the repeat visit patterns underlying these two groups are associated with an increase in total time spent away. It also suggests a higher degree of residential ambiguity especially for the Floaters. Since members of this group report visits to at least two places in addition to the interview household, it is unclear whether

those away more than half the time are spending a majority of time at any one place. If time spent at each place is roughly equal, it is easy to imagine Floaters not being rostered at any of them or at more than one of them. Conversely, by definition we can assume the Boomerangs who were away more than half the reference period spent the majority of their time at the only other place they reported visiting. Assuming time plays a role in defining a sense of household membership, then presumably, the Boomerangs have a better chance of being counted because the majority of their time is being spent at the other place.

5. USUAL RESIDENCE AND MOBILITY

We next explored whether temporary mobility has an impact on the household respondent's determination of a person as a "usual resident". On the 1990 U.S. census form, respondents were instructed to list persons at the place where the person lives or sleeps most of the time. The LSS asked household respondents whether they considered the interview household to be the "usual residence, that is the place where [you/NAME] live(s) and sleep(s) most of the time". They were also asked to report whether "[you/NAME] have a usual residence somewhere else?" While this method is not a perfect replication of a census roster it provides an approximation of who, out of all those rostered during the LSS, the householder might naturally have included or excluded on a census form or current survey.

Table 4 presents a cross-classification of usual residence assignment by mobility status. A combination of the usual residence questions resulted in four classification possibilities: usual residence at the interview household only, usual residence at someplace other than the interview household only, usual residence at both the interview household and another place, and usual residence at no place. (The category of "no place" was extremely small (less than 1%) and was combined into the category of "other place"). Assuming that answers of "other place" equate to being left off the census form, we see that overall, only around 4% of persons with a greater-than-casual association to the interview households might have been left off. Overall, the distribution of usual resident classifications significantly differed according to mobility type.

As might be expected, nearly all of the persons who spent every night at the interview household during the reference period were considered usual residents there (rounded to 100%). The most obvious deviation among categories is noticeable for the Boomerangs and Floaters. Between 20-25% of the people in these two groups were characterized by household respondents as usual residents someplace other than the interview household. This looks very different from both the 1-shots and No-repeat groups, where only 2% and 5%, respectively, were considered usual

residents someplace else. These results suggest that the latter two groups typify mobility associated with pleasure or business but for persons with a firm tie to the household while the Boomerangs and the Floaters are more likely to include persons with a less-established association to the household. For this reason, and the fact that a sizable percentage of the HTE population were found in these two categories, the Boomerangs and Floaters arguably have the more interesting coverage implications and raise several questions. For example, do these persons get counted at one place, all places or no place? Additionally, where should they be counted?

Table 4

Where Does Household Respondent Consider Person to be a "Usual Resident"? (Weighted % and standard errors)

Where Usual Resident ?	Non Mobile	1-Shots	Boomerangs	No Repeat	Floaters	Total
Interview HH Only	100% (0.2)	97% (2.0)	71% (12.1)	95% (4.2)	70% (10.0)	95% (1.7)
Some Other Place	0% (-)	2% (1.8)	25% (11.0)	5% (4.2)	20% (9.4)	4% (1.5)
Both Places	0% (-)	1% (0.4)	4% (2.1)	0% (-)	10% (7.3)	1% (0.5)
Unweighted N	716	314	186	101	134	1,451

Jackknife chi-square = 2.79, $p < .05$, $d.f. = 8$

That a relatively large percentage of the Boomerangs and Floaters are considered residents some place other than the interview household suggests the potential for undercounting. On the other hand, 10% of the Floaters are defined as usual residents at both the interview household and another place suggests potential for overcoverage. The weighted number of Boomerangs and Floaters in these uncertain residency situations (usual residents elsewhere or at both places) represent approximately 4% of the total population. From this more global perspective, it seems that a non-trivial segment of the population is at risk of some type of coverage error.

6. MODELING OF USUAL RESIDENCE AND MOBILITY

Our final section statistically models the household respondent's determination of usual residence. This analysis goes beyond the descriptive findings of the typology to explore whether mobility impacts the householder's conceptualization of residence. The assignment of usual residence by the householder served as the dependent variable in a series of models. The dependent variable consisted of two categories: 1) usual resident at the interview household and 2) not a usual resident at the interview household. Persons considered to have a usual residence at both the interview household and another place were put

into the first category. Predictor variables included age, sex, race, time away, and the mobility typology. The final models reported in Table 5, all of which include terms for the interaction of the independent variables, are equivalent to logit models for usual residence.

The first model tested mobility as a dichotomous measure: those with no mobility (the Non-mobile) and those having spent at least one night away from the interview household (the 1-shot, No-Repeat, Boomerang and Floater categories combined). This model established first whether temporary mobility was a significant predictor of residency status regardless of the mobility pattern exhibited. This "first-cut" was necessary because approximately 50% of the sample fell into the Non-mobile category and second, because the Non-mobile group was extremely skewed toward the usual resident category of the dependent variable. Consequently, models that attempted to include all five categories of the mobility typology were misspecified due to a large number of zero fitted cells.

Results from the model with the dichotomous mobility measure and sex yielded a relatively good "fit" of the data (Jackknife X^2 for overall goodness of fit = .28, $d.f.$ = 2, p = .27. Neither race nor age improved the fit. Parameter estimates indicated that persons in the Non-mobile category were more likely to be classified as usual residents than those having some mobility (not shown).

Having established that mobility was significantly related to residency status, we next explored whether the pattern of temporary mobility was a predictor. First, we tested an independence baseline model to predict usual residence (U). The predictors consisted of a mobility variable (M), sex (S), and the amount of time spent away (T). The mobility variable was comprised of the four mobile categories (1-Shots, No-Repeats, Boomerangs, and Floaters). Amount of time spent away was split into two categories: less than half the reference period and half or more of the reference period. Race and age were excluded since neither improved the fit of the data.

Table 5
Goodness-of-Fit Tests and Parameter Estimates for Log-Linear Models of the Effect of Sex (S), Temporary Mobility (M), and Length of Time Away (T) on Determination of Usual Residence Status (U)

A. Goodness of Fit Test			
Model	<i>d.f.</i>	(U) Usual Residence Status Chi-square ⁺	<i>p</i>
1. U, SMT	15	4.79	.00
2. US, UM, UT, SMT	10	1.06	.12
3. UTM, USM, SMT	4	0.78	.16
B. Parameter Estimates, Model 3			
	beta	s.e.	std. value
(M) MOBILITY:			
1-Shots	1.08	.40	2.71*
Boomerangs	-1.54	.39	-3.94*
No-Repeats	.83	.58	1.43
Floaters	-.38	.47	-.80
(S) SEX:			
(Males)	.39	.27	1.44
(T) TIME AWAY:			
(> ½ ref. period)	-1.78	.27	-6.52*
(U)*(S)*(M) INTERACTION (Males)			
1-Shots	-.64	.43	-1.48
Boomerangs	.69	.58	1.18
No-Repeats	.85	.62	1.37
Floaters	-.90	.42	-2.14*
(U)*(M)*(T) INTERACTION (> ½ ref. period)			
1-Shots	-.72	.48	-1.50
Boomerangs	-1.20	.54	-2.26*
No-Repeats	1.57	.74	2.12*
Floaters	.36	.41	0.88

⁺ Jackknife Pearson chi-square for overall fit.

* Significant at the .05 level.

The baseline model (U, SMT) did not fit the data well so we rejected the null hypothesis that assignment of usual residence is independent of mobility pattern, sex, and amount of time spent away (Jackknife X^2 overall goodness of fit = 4.79, $d.f.$ = 15, p = .00, see Table 5). We then fitted a main effects model (2) which includes the additive effects of S, M and T upon U (US, UM, UT, SMT). This model yielded a good fit (Jackknife X^2 overall goodness of fit = 1.06, $d.f.$ = 10, p = .12). Lastly, a model (3) including two interaction terms was also fitted (UTM, USM, SMT). This model assumes interactive effects of T*M and of S*M on U. A comparison between the main effects and interaction model suggested that several interactions were significant and should be retained (comparison Jackknife X^2 = 1.99, $d.f.$ = 6, p = .02). Table 5 contains the overall goodness of fit tests along with the parameter estimates from the best fitting interaction model (UTM, USM, SMT – Jackknife X^2 overall goodness of fit = 0.78, $d.f.$ = 4, p = .16.)

The parameter estimates from Table 5 illustrate that temporary mobility has a significant main effect on assignment of usual residence in model 3 which controls for sex, amount of time spent away, and several interactions. Two of the mobility categories had significant beta coefficients albeit the directions were opposite. The 1-Shots were significantly more likely to be defined as usual residents (b = +1.08). Conversely, the Boomerangs had a negative parameter estimate (b = -1.54) meaning that the odds of being defined a usual resident were significantly decreased for this group.

Time spent away from the interview household had by far the largest effect on predicting usual residence with a strong negative association (b = -1.78). This means that for our temporarily mobile population, those away half or more of the reference period were significantly less likely to be considered usual residents than those away less than half of the time. Sex did not have a significant main effect, but was involved in a significant interaction. The interaction appears in the Floater group where male Floaters were less likely to be categorized as usual residents than female Floaters (b = -.90). Further investigation revealed few clues to explain this finding. Male and female Floaters differed little in the types of places they visited, their reasons for visiting, and the relation to the householder of places they visited (relative versus non-relative). Perhaps the interaction reflects differences in other social attachments such as presence of children, personal belongings, and/or contribution of resources.

The bottom of table 5 indicates that the interaction between usual residence, mobility and amount of time spent away is rather complex. The amount of time spent away appears to affect usual residence status for some types of mobility but not for others. The interaction coefficient is significant and negative for the Boomerangs (b = -1.20). Thus, the odds of being defined a usual resident are even lower for Boomerangs away half or more of the reference period compared to other groups away for a similar amount

of time. This suggests that persons who “boomerang” back and forth between two households will be considered usual residents at the place they spend the majority of time.

However, for the No-repeats, the coefficient is significant and *positive*, essentially canceling out time away’s negative main effect ($1.57 + -1.78 = -0.21$). For this group, the amount of time spent away appears to have no association with usual residence assignment. Apparently, factors other than time may be more important in the cognitive process of determining where these persons “reside.” One hypothesis is that No-repeaters are persons who must travel for a living and who, despite their frequent mobility and long periods away, clearly “belong” to a stable residence. This notion supports findings from a vignette study that found respondents did not require a stated rule to be able to correctly identify the usual residence of persons described as being away on business travel. Such persons were “intuitively” perceived to be part of the households from which they were away (Gerber, Wellens and Keeley 1996).

7. CONCLUSIONS

Temporary mobility, as defined in our research, involves long and short, frequent and infrequent, patterned and unpatterned movement away from, but often back to, a single residence. Such mobility has long been hypothesized to contribute toward census and survey coverage error by blurring the concept of who exactly lives or stays at a particular household.

Our sample of persons having a more-than-casual association to households indicated a fair amount of temporary mobility over a two-three month period. Interesting demographic differences were noted in the level of mobility as well as the pattern of mobility reported. The “hard to enumerate” (HTE) group (black/Hispanic males between 18 and 29) were found to cluster in the Boomerang and Floater groups, suggesting a repeat pattern of temporary mobility. We suspect these groups include persons having strong attachments to multiple households, for example an adult son who splits time between a parent and girlfriend’s or a young mother who stays periodically at different kin-network households to receive assistance with child care.

Besides the inclusion of the types of persons who tend to be missed in censuses and surveys, other considerations point to the Boomerang and Floaters as being of particular interest. First, compared to the other mobility categories, these groups spent a longer time away from the households in which they were “found” and second, were more often classified as having a usual residence someplace other than the household in which they were found. It is difficult to estimate how much this type of mobility contributes toward undercounting. However, it is very noteworthy that half the HTE population fall in either the Boomerang or Floater group. It seems more than a coincidence that such a large segment of this population belong to one of the two mobility groups most easily labeled “residentially ambiguous.”

The log-linear analysis suggests that there is not a clearcut, simple relationship between temporary mobility and assignment of usual residence. We do not find that the greater the amount of temporary mobility the less the chance of being defined a usual resident. Instead, the relationship seems more driven by the pattern of movement. For example, the traveling salesman or truck driver who reports the greatest variety of places visited and the largest number of visits may, nonetheless, have less residential ambiguity than a person visiting only one other place but making many repeat visits. And, in fact, this proved to be the case for the No-Repeats for whom the amount of time spent away had essentially no relation to usual residence assignment.

Our exploration of temporary mobility represents a new research direction for the study of within-household census and survey coverage error. Two recommendations for improving census and survey coverage are offered. First, survey organizations should explore the possibility of directly measuring the association between temporary mobility and incidents of census and survey undercoverage. This could be accomplished by adding questions about mobility to post-census coverage interviews used to estimate the number of people missed or counted in error. If the correlation between coverage error and mobility is significant, then survey methods and procedures could be adjusted to try and reduce it. For example, new roster probes could be added to census forms and nonresponse follow-up interviews, the aim being to find more of the Boomerangs and Floaters. Measures of temporary mobility might also prove to be a powerful predictor variable when statistically modeling the undercount. While admittedly in the early stages, temporary mobility looks promising as an avenue to better understanding household coverage error.

ACKNOWLEDGMENTS

The authors wish to thank Elizabeth A. Martin, Theresa J. DeMaio, Robert E. Fay and three anonymous reviewers for insightful comments on earlier versions of this paper. This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau publications. This report is released to inform interested parties of research and to encourage discussion.

REFERENCES

- CHAKRABARTY, R. (1992). Coverage of the Current Population Survey (CPS) Relative to the 1990 Census. Unpublished U.S. Bureau of the Census memorandum to the record, February 20, 1992.
- DE LA PUENTE, M. (1993). Why are people missed or erroneously included by the census: a summary of findings from ethnographic coverage reports. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 29-66.
- DRESSLER, W., HOEPPNER, S., and PITTS, B. (1985). Household structure in a southern black community. *American Anthropologist*, 87, 835-862.
- ELLIS, Y. (1994). Categorical Data Analysis of Census Omissions, Internal Memorandum, Washington D.C: U.S. Bureau of the Census.
- ELLIS, Y. (1993). Census Error Study. U.S. Bureau of the Census, 1990 Preliminary Research and Evaluation Memorandum No. 248.
- FAY, R.E. (1989a). An analysis of within-household undercoverage in the current population survey. *Proceedings of the 1989 Annual Research Conference*, U.S. Bureau of the Census, 156-175.
- FAY, R.E. (1989b). CPLX: Contingency Table Analysis for Complex Sample Designs. Program Documentation. Unpublished document, U.S. Bureau of the Census.
- FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FEIN, D.J., and WEST, K. (1988). Toward a theory of coverage error: an exploration of data from the 1986 Los Angeles test census. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 540-562.
- GERBER, E., WELLENS, T., and KEELEY, C. (1996). Who Lives Here?: The Use of Vignettes in Household Roster Research. Paper presented at the annual meeting of the American Association for Public Opinion Research, Salt Lake City.
- GERBER, E. (1994). The Language of Residence: Respondent Understandings and Census Rules. Unpublished report of the Cognitive Study of Living Situations. Center for Survey Methods Research, U.S. Bureau of the Census.
- GRIFFIN, D., and MORIARTY, C. (1992). Characteristics of Census Errors. U.S. Bureau of the Census, 1990 Preliminary Research and Evaluation Memorandum No. 179.
- HAINER, P. (1987). A Brief and Qualitative Anthropological Study Exploring the Reasons for Census Coverage Error Among Low Income Black Households. Report for the Census for Survey Methods Research, U.S. Bureau of the Census, April 8, 1987.
- HAINER, P., HINES, C., MARTIN, E.A., and SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 513-539.
- HANSEN, K. (1994). Geographical Mobility: March 1993 to March 1994. Current Population Reports, Population Characteristics P20-485. U.S. Department of Commerce.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.
- HUDGINS, J.L., and HOLMES, B.J. (1993). The impact of social and economic marginality on the underenumeration of African American males. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 153-166.
- LIPTON, S.G., and ESTRADA, L.F. (1993). Factors associated with undercount rates in Los Angeles county. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 83-102.
- LYNCH, J.T., WITT, M., BRANSON, S., and ARDINI, M. (1993). Living Situation Survey: Final Methods Report, Unpublished report, Research Triangle Institute: Research Triangle Park.

- MAHLER, S. (1993). Alternative Enumeration of Undocumented Salvadorans on Long Island. Prepared under Joint Statistical Agreement 89-46 with Columbia University. U.S. Bureau of the Census, Washington, D.C.
- MARTIN, J.M., and MARTIN, E.P. (1985). *The Helping Tradition in the Black Family and Community*. Silver Spring, Md.: National Association of Social Workers.
- MORIARTY, C.L., and CHILDERS, D. (1993). Analysis of Census Omissions: Preliminary Results. U.S. Bureau of the Census, DSSD 1990 REX Memorandum Series #PP-8.
- MONTOYA, D. (1992). Ethnographic Evaluation of the Behavioral Causes of Undercount: Woodburn, Oregon. Ethnographic Evaluation of the 1990 Decennial Census Report #10. Prepared under Joint Statistical Agreement 89-30 with the University of Oklahoma. U.S. Bureau of the Census: Washington, D.C.
- PENNIE, K. (1990). Coverage Comparisons Between the 1990 Census and Current Population Survey (CPS). Unpublished U.S. Bureau of the Census memorandum to Preston Jay Waite.
- ROBINSON, J.G, AHMED, B., DAS GUPTA, P., and WOODROW, K.A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1071.
- ROMERO, M. (1992). Ethnographic Evaluation of the Behavioral Causes of Census Undercount of Undocumented Immigrants and Salvadorans in the Mission District of San Francisco, California. Ethnographic Evaluation of the 1990 Decennial Census, Report #18. Prepared under Joint Statistical Agreement 89-41 with the San Francisco State University Foundation. U.S. Bureau of the Census, Washington, D.C.
- SHAPIRO, G., DIFFENDAL, G., and CANTOR, D. (1993). Survey undercoverage: major causes and new estimates of magnitude. *Proceedings of the 1993 Annual Research Conference*, U.S. Bureau of the Census, 638-663.
- STACK, C.B. (1974). *All Our Kin: Strategies for Survival in the Black Community*. New York: Harper and Row.
- SWEET, E.M. (1994). Roster research results from the living situation survey. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 415-433.
- U.S. DEPARTMENT OF LABOR (1994). *Employment and Earnings*. Bureau of Labor Statistics, January 1994: Washington, D.C.
- VALENTINE, C., and VALENTINE, B. (1971). *Missing Men: A Comparative Methodology Study of Underenumeration and Related Problems*. Report to the U.S. Bureau of the Census, May 3, 1971.
- VELASCO, A. (1992). Ethnographic Evaluation of the Behavioral Causes of Undercount in the Community of Sherman Heights, California. Ethnographic Evaluation of the 1990 Decennial Census, Report #22. Prepared under Joint Statistical Agreement 89-42 with the Chicano Federation of San Diego County. U.S. Bureau of the Census, Washington, D.C.
- WINGERD, J. (1992). Urban Haitians: Documented/Undocumented in a Mixed Neighborhood. Ethnographic Evaluation of the 1990 Decennial Census, Report #7. Prepared under Joint Statistical Agreement # 90-10 with the Community Service Council of Broward County, Inc. U.S. Bureau of the Census, Washington, D.C.

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents

Volume 14, Number 1, 1998

Data Collection Mode Effects on Responses to Attitudinal Questions <i>Peter Lynn</i>	1
Effects of Interview Mode on Measuring Depression in Younger Adults <i>William S. Aquilino</i>	15
Space/Time Variations and rolling Samples <i>Leslie Kish</i>	31
Examining the Revisions in Monthly Retail and Wholesale Trade Surveys Under a Rotating Panel Design <i>Patrick J. Cantwell and Carol V. Caldwell</i>	47
Random-Effects Models for Smoothing Poststratification Weights <i>Laura C. Lazzeroni and Roderick J.A. Little</i>	61
Estimation of Identification Disclosure Risk in Microdata <i>Guang Chen and Sallie Keller-McNulty</i>	79
Book and Software Reviews	97
In Other Journals	113

All inquires about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S-104 51 Stockholm, Sweden.

CONTENTS

TABLE DES MATIÈRES

Volume 26, No. 1, March/mars 1998

Editorial/Éditorial

Gareth O. ROBERTS and Jeffrey S. ROSENTHAL

Markov-chain Monte Carlo: Some practical implications of theoretical results

Discussion: Hemant ISHWARAN

Discussion: Neal MADRAS

Rejoinder: Gareth O. ROBERTS and Jeffrey S. ROSENTHAL

Helen ASLANIDOU, Dipak K. DEY and Debajyoti SINHA

Bayesian analysis of multivariate survival data using Monte Carlo methods

Laura VENTURA

Higher-order approximations for Pitman estimators and for optimal compromise estimators

Richard DYKSTRA, Subhash KOCHAR and Tim ROBERTSON

Restricted tests for testing independence of time to failure and cause of failure in a competing-risks model

Chul Gyu PARK, Chu-In Charles LEE and Tim ROBERTSON

Goodness-of-fit test for uniform stochastic ordering among several distributions

S.R. PAUL and A.S. ISLAM

Joint estimation of the mean and dispersion parameters in the analysis of proportions: a comparison of efficiency and bias

Y. LEE and J.A. NELDER

Generalized linear models for the analysis of quality-improvement experiments

T. Rolf TURNER, Murray A. CAMERON and Peter J. THOMSON

Hidden Markov chains in generalized linear models

Markus ABT and William J. WELCH

Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes

A.C. DAVISON and J.E. STAFFORD

The score function and a comparison of various adjustments of the profile likelihood

Arthur B. YEH

A bootstrap procedure in linear regression with non-stationary errors

K. VIRASWAMI and N. REID

A note on the likelihood-ratio statistic under model misspecification

Brajendra C. SUTRADHAR and Zhende QU

On approximate likelihood inference in Poisson mixed model

Kilani GHOUDI, Abdelhaq KHOUDRAJI et Louis-Paul RIVEST

Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles

CONTENTS

TABLE DES MATIÈRES

Volume 26, No. 2, June/juin 1998

D.J. DUPUIS and C.A. FIELD

Robust estimation of extremes

Ömer ÖZTÜRK and Thomas P. HETTMANSPERGER

Simultaneous robust estimation of location and scale parameters: A minimum-distance approach

Douglas P. WIENS, Eden K.H. WU and Julie ZHOU

On the trimmed mean and minimax-variance L-estimation in Kolmogorov neighbourhoods

Ana BIANCO and Graciela BOENTE

Robust kernel estimators for additive models with dependent observations

Eden K.H. WU and Julie ZHOU

Efficient bias-robust M-estimators of location in Kolmogorov normal neighborhoods

Nadia BENSÂÏD et Jean-Pierre FABRE

Convergence de l'estimateur à noyau de dérivées de Radon-Nikodym générales dans le cas mélangeant

Pietro MULIERE and Lucas TARDELLA

Approximating distributions of random functions of Ferguson-Dirichlet priors

C. Andy TSAO and J.T. Gene HWANG

Improved confidence estimators for Fieller's confidence sets

Germai CHEN

The run length distributions of the R , s and s^2 charts when σ is estimated

Dongchu SUN, Malay GHOSH and Asit P. BASU

Bayesian analysis for a stress-strength system under noninformative priors

José BERRENDERO, Sonia MAZZI, Juan ROMO and Ruben ZAMAR

On the explosion rate of maximum-bias functions...

Ajit C. TAMHANE, Wei LIU and Charles W. DUNNETT

A generalized step-up-down multiple test procedure

Alain BERLINET, Ali GANNOUN et Eric MATZNER-LØBER

Normalité asymptotique d'estimateurs convergents du mode conditionnel

First Announcement

Call for Papers

IASS SATELLITE CONFERENCE ON SMALL AREA ESTIMATION

Riga, Latvia, 20-21 August 1999

The Satellite Conference on Small Area Estimation will follow the ISI session in Helsinki. It is intended to cover aspects of both theoretical background in small area estimation and practical application of different estimation methods for small area statistics. This includes sample design for small area statistics (national experiences), new developments in the field of estimation for small area statistics and successful applications of small area estimation techniques, including those that use data from administrative systems. Small area statistics is a subject of great interest in many countries. Several statistical agencies in Western countries have introduced vigorous programmes to meet this new demand, with a view toward producing efficient and high quality statistics. Several international conferences and seminars have been organised in the last years and others are yet to be organised. Furthermore, significant research on both the theoretical and practical aspects of small area estimation is conducted at various universities and some national statistical offices.

The Conference is organised on the initiative of the Baltic countries, and is aimed at improving knowledge transfer of new methods. The proceedings of the Conference should be of interest to all statisticians working in this field but it is of particular interest for the economies in transition in Central and Eastern European countries and the former Soviet Union countries, where complete reporting and complete statistical investigations are to be replaced or have been replaced with sample surveys, the production of reliable small area statistics has emerged as a pressing and frequently difficult and costly problem.

The conference proceedings will be opened by Dr. Danny Pfeffermann, who will provide an overview of the New Developments in Small Area Estimation. Initial plans also include holding a one day Short Course on Small Area Estimation immediately preceding the Conference, in order to allow some participants to acquire the basic knowledge that would allow them to appreciate fully the proceedings of the conference. The meeting is sponsored by the International Association of Survey Statisticians (IASS), the Central Statistical Bureau of Latvia (CSBL), and the University of Latvia (UL).

The members of the International Programme Committee are: Ödon Éltető (Hungary), Wayne A. Fuller (USA), Jan Kordos (Poland, chair), John Kovar (Canada), Juris Krumins (Latvia), Janis Lapinš (Latvia), Danny Pfeffermann (Israel), Richard Platek (Canada), J.N.K. RAO (Canada), Carl-Erik Särndal (Canada), Dennis Trewin (Australia) and Janusz Wywiał (Poland).

Abstracts of proposed papers should include full information on authors and their affiliations, and the contact address (including e-mail and fax) and a text of 200-300 words. The deadline for submission is December 31, 1998. Earlier submissions are encouraged and notifications of acceptance will be sent out as soon as possible. Acceptance is conditional on the attendance of the meeting by at least one of the authors. Abstract should be submitted, preferably via e-mail (in ASCII or WORD 6.0), or by fax or by mail to:

Jan Kordos, Al. Niepodległości 208, 00-925 Warsaw, Poland;
Fax: (0048-22) 825-03-95; E-mail: kordos@gus.stat.gov.pl
or to any other members of the Programme Committee

It is the intention of the Programme Committee to publish the papers presented at the Conference in a special Proceedings of the Conference issue. The papers may also be published in any journal after the Conference.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue of *Survey Methodology* (Vol. 19, No. 1 and onward) as a guide and note particularly the points below. Accepted articles must be submitted in machine-readable form, preferably in WordPerfect. Other word processors are acceptable, but these also require paper copies for formulas and figures.

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points ci-dessous. Les articles acceptés doivent être soumis sous forme de fichiers de traitement de texte, préféablement WordPerfect. Les autres logiciels sont acceptables, mais une version sur papier sera alors exigée pour le traitement des formules et des figures.

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; l, I).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
Exemple: Cochran (1977, p. 164).
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

First Announcement Call for Papers

IASS SATELLITE CONFERENCE ON SMALL AREA ESTIMATION Riga, Latvia, 20-21 August 1999

The Satellite Conference on Small Area Estimation will follow the ISI session in Helsinki. It is intended to cover aspects of both theoretical background in small area estimation and practical application of different estimation methods for small area statistics. This includes sample design for small area statistics (national experiences), new developments in the field of estimation for small area statistics and successful applications of small area estimation techniques, including those that use data from administrative systems. Small area statistics is a subject of great interest in many countries. Several statistical agencies in Western countries have introduced vigorous programmes to meet this new demand, with a view toward producing efficient and high quality statistics. Several international conferences and seminars have been organised in the last years and others are yet to be organised. Furthermore, significant research on both the theoretical and practical aspects of small area estimation is conducted at various universities and some national statistical offices.

The Conference is organised on the initiative of the Baltic countries, and is aimed at improving knowledge transfer of new methods. The proceedings of the Conference should be of interest to all statisticians working in this field but it is of particular interest for the economies in transition in Central and Eastern European countries and the former Soviet Union countries, where complete reporting and complete statistical investigations are to be replaced or have been replaced with sample surveys, the production of reliable small area statistics has emerged as a pressing and frequently difficult and costly problem.

The conference proceedings will be opened by Dr. Danny Pfeffermann, who will provide an overview of the New Developments in Small Area Estimation. Initial plans also include holding a one day Short Course on Small Area Estimation immediately preceding the Conference, in order to allow some participants to acquire the basic knowledge that would allow them to appreciate fully the proceedings of the conference. The meeting is sponsored by the International Association of Survey Statisticians (IASS), the Central Statistical Bureau of Latvia (CSBL), and the University of Latvia (UL).

The members of the International Programme Committee are: Odon Elieć (Hungary), Wayne A. Fuller (USA), Jan Kordos (Poland, chair), John Kovar (Canada), Juris Kruminš (Latvia), Janis Lapins (Latvia), Danny Pfeffermann (Israel), Richard Platek (Canada), J.N.K. RAO (Canada), Carl-Erik Särndal (Canada), Dennis Trewin (Australia) and Janusz Wywiał (Poland).

Abstracts of proposed papers should include full information on authors and their affiliations, and the contact address (including e-mail and fax) and a text of 200-300 words. The deadline for submission is December 31, 1998. Earlier submissions are encouraged and notifications of acceptance will be sent out as soon as possible. Acceptance is conditional on the attendance of the meeting by at least one of the authors. Abstract should be submitted, preferably via e-mail (in ASCII or WORD 6.0), or by fax or by mail to:

Jan Kordos, Al. Niepodległości 208, 00-925 Warsaw, Poland;
Fax: (0048-22) 825-03-95; E-mail: kordos@gus.stat.gov.pl
or to any other members of the Programme Committee

It is the intention of the Programme Committee to publish the papers presented at the Conference in a special Proceedings of the Conference issue. The papers may also be published in any journal after the Conference.

Volume 26, No. 2, June/juin 1998

D.J. DUPUIS and C.A. FIELD
Robust estimation of extremes

Ömer ÖZTÜRK and Thomas P. HETTMANSPERGER
Simultaneous robust estimation of location and scale parameters: A minimum-distance approach

Douglas P. WIENS, Eden K.H. WU and Julie ZHOU
On the trimmed mean and minimax-variance L-estimation in Kolmogorov neighbourhoods

Ana BIANCO and Graciela BOENTE
Robust kernel estimators for additive models with dependent observations

Eden K.H. WU and Julie ZHOU
Efficient bias-robust M-estimators of location in Kolmogorov normal neighbourhoods

Nadia BENSADID et Jean-Pierre FABRE
Convergence de l'estimateur à noyau de dérivées de Radon-Nikodym générales dans le cas mélangé

Pietro MULIERE and Lucas TARDELLA
Approximating distributions of random functions of Ferguson-Dirichlet priors

C. Andy TSAO and J.T. Gene HWANG
Improved confidence estimators for Fieller's confidence sets

Gernai CHEN
The run length distributions of the R , s and s^2 charts when σ is estimated

Dongchu SUN, Malay GHOSH and Asit P. BASU
Bayesian analysis for a stress-strength system under noninformative priors

José BERRENDERO, Sonia MAZZI, Juan ROMO and Ruben ZAMAR
On the explosion rate of maximum-bias functions...

Ajit C. TAMHANE, Wei LIU and Charles W. DUNNETT
A generalized step-up-down multiple test procedure

Alain BERLINET, Ali GANNOUN et Eric MATZNER-LØBER
Normalité asymptotique d'estimateurs convergents du mode conditionnel

Volume 26, No. 1, March/mars 1998

Editorial/Editorial

Gareth O. ROBERTS and Jeffrey S. ROSENTHAL

Markov-chain Monte Carlo: Some practical implications of theoretical results

Discussion: Hemant ISHWARAN

Discussion: Neal MADRAS

Rejoinder: Gareth O. ROBERTS and Jeffrey S. ROSENTHAL

Helen ASLANIDOU, Dipak K. DEY and Debajyoti SINHA

Bayesian analysis of multivariate survival data using Monte Carlo methods

Laura VENTURA

Higher-order approximations for Pitman estimators and for optimal compromise estimators

Richard DYKSTRA, Subhash KOCHAR and Tim ROBERTSON

Restricted tests for testing independence of time to failure and cause of failure in a competing-risks model

Chul Gyu PARK, Chu-In Charles LEE and Tim ROBERTSON

Goodness-of-fit test for uniform stochastic ordering among several distributions

S.R. PAUL and A.S. ISLAM

Joint estimation of the mean and dispersion parameters in the analysis of proportions: a comparison of efficiency and bias

Y. LEE and J.A. NELDER

Generalized linear models for the analysis of quality-improvement experiments

T. Rolf TURNER, Murray A. CAMERON and Peter J. THOMSON

Hidden Markov chains in generalized linear models

Markus ABT and William J. WELCH

Fisher information and maximum likelihood estimation of covariance parameters in Gaussian stochastic processes

A.C. DAVISON and J.E. STAFFORD

The score function and a comparison of various adjustments of the profile likelihood

Arthur B. YEH

A bootstrap procedure in linear regression with non-stationary errors

K. VIRASWAMI and N. REID

A note on the likelihood-ratio statistic under model misspecification

Brayendra C. SUTRADHAR and Zhende QU

On approximate likelihood inference in Poisson mixed model

Kilani CHOUDI, Abdelhak KHOUDRAJI et Louis-Paul RIVEST

Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 14, Number 1, 1998

Data Collection Mode Effects on Responses to Attitudinal Questions <i>Peter Lynn</i>	1
Effects of Interview Mode on Measuring Depression in Younger Adults <i>William S. Aquilino</i>	15

Space/Time Variations and rolling Samples <i>Leslie Kish</i>	31
Examining the Revisions in Monthly Retail and Wholesale Trade Surveys Under a Rotating Panel Design <i>Patrick J. Cantwell and Carol V. Caldwell</i>	47

Random-Effects Models for Smoothing Poststratification Weights <i>Laura C. Lazzeroni and Roderick J.A. Little</i>	61
Estimation of Identification Disclosure Risk in Microdata <i>Guang Chen and Sallie Keller-McNulty</i>	79

Book and Software Reviews	97
In Other Journals	113

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S-104 51 Stockholm, Sweden.

VELASCO, A. (1992). Ethnographic Evaluation of the Behavioral Causes of Undercount in the Community of Sherman Heights, California. Ethnographic Evaluation of the 1990 Decennial Census, Report #22. Préparé sous le Joint Statistical Agreement 89-42 avec Chicano Federation of San Diego County. U.S. Bureau of the Census, Washington, D.C.

WINGBERD, J. (1992). Urban Haitians: Documented/Undocumented in a Mixed Neighborhood. Ethnographic Evaluation of the 1990 Decennial Census, Report #7. Préparé sous le Joint Statistical Agreement # 90-10 avec Community Service Council of Broward County, Inc. U.S. Bureau of the Census, Washington, D.C.

BIBLIOGRAPHIE

- CHAKRABARTY, R. (1992). Coverage of the Current Population Survey (CPS) Relative to the 1990 Census. U.S. Bureau of the Census note de service non-publiée, le 20 février 1992.
- DE LA PUENTE, M. (1993). Why are people missed or erroneously included by the census: a summary of findings from ethnographic coverage reports. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 29-66.
- DRESSLER, W., HOEPFNER, S., et PITTS, B. (1985). Household structure in a southern black community. *American Anthropologist*, 87, 835-862.
- ELLIS, Y. (1994). Categorical Data Analysis of Census Omissions, Note de service interne, Washington D.C: U.S. Bureau of the Census.
- ELLIS, Y. (1993). Census Error Study. U.S. Bureau of the Census, 1990 Preliminary Research and Evaluation Memorandum No. 248.
- FAY, R.E. (1989a). An analysis of within-household undercoverage in the current population survey. *Proceedings of the 1989 Annual Research Conference*, U.S. Bureau of the Census, 156-175.
- FAY, R.E. (1989b). CPLX: Contingency Table Analysis for Complex Sample Designs. Program Documentation. Article non publié, U.S. Bureau of the Census.
- FAY, R.E. (1985). A jackknifed chi-squared test for complex samples. *Journal of the American Statistical Association*, 80, 148-157.
- FEIN, D.J., et WEST, K. (1988). Toward a theory of coverage error: an exploration of data from the 1986 Los Angeles test census. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 540-562.
- GERBER, E., WELLES, T., et KEELEY, C. (1996). Who Lives Here?: The Use of Vignettes in Household Roster Research. Article présenté à la réunion annuelle de l' American Association for Public Opinion Research, Salt Lake City.
- GERBER, E. (1994). The Language of Residence: Respondent Understandings and Census Rules. Rapport non publié du Cognitive Study of Living Situations. Center for Survey Methods Research, U.S. Bureau of the Census.
- GRIFFIN, D., et MORIARTY, C. (1992). Characteristics of Census Errors. U.S. Bureau of the Census, 1990 Preliminary Research and Evaluation Memorandum No. 179.
- HAINER, P. (1987). A Brief and Qualitative Anthropological Study Exploring the Reasons for Census Coverage Error Among Low Income Black Households. Rapport pour le Census for Survey Methods Research, U.S. Bureau of the Census, 8 avril 1987.
- HAINER, P., HINES, C., MARTIN, E.A., et SHAPIRO, G. (1988). Research on improving coverage in household surveys. *Proceedings of the Fourth Annual Research Conference*, U.S. Bureau of the Census, 513-539.
- HANSEN, K. (1994). Geographical Mobility: March 1993 to March 1994. Current Population Reports, Population Characteristics P20-485. U.S. Department of Commerce.
- HOGAN, H. (1993). The 1990 post-enumeration survey: operations and results. *Journal of the American Statistical Association*, 88, 1047-1060.
- HUDGINS, J.L., et HOLMES, B.J. (1993). The impact of social and economic marginality on the underenumeration of African American males. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 153-166.
- LIPTON, S.G., et ESTRADA, L.F. (1993). Factors associated with undercount rates in Los Angeles county. *Proceedings of the 1993 Research Conference on Undercounted Ethnic Populations*, 83-102.
- LYNCH, J.T., WITT, M., BRANSON, S., et ARDINI, M. (1993). Living Situation Survey: Final Methods Report, Rapport non publié, Research Triangle Institute; Research Triangle Park.
- MAHLER, S. (1993). Alternative Enumeration of Undercounted Salvadorans on Long Island. Prépare sous le Joint Statistical Agreement 89-46 avec Columbia University. U.S. Bureau of the Census, Washington, D.C.
- MARTIN, J.M., et MARTIN, E.P. (1985). The Helping Tradition in the Black Family and Community. Silver Spring, Md.: National Association of Social Workers.
- MORIARTY, C.L., et CHILDERS, D. (1993). Analysis of Census Omissions: Preliminary Results. U.S. Bureau of the Census, DSSD 1990 REX Memorandum Series #PP-8.
- MONTROY, D. (1992). Ethnographic Evaluation of the Behavioral Causes of Undercount: Woodburn, Oregon. Ethnographic Evaluation of the 1990 Decennial Census Report #10. Prépare sous le Joint Statistical Agreement 89-30 avec University of Oklahoma. U.S. Bureau of the Census; Washington, D.C.
- PENNIE, K. (1990). Coverage Comparisons Between the 1990 Census and Current Population Survey (CPS). U.S. Bureau of the Census note de service non-publiée à Preston Jay Waite.
- ROBINSON, J.G, AHMED, B., DAS GUPTA, P., et WOODROW, K.A. (1993). Estimation of population coverage in the 1990 United States census based on demographic analysis. *Journal of the American Statistical Association*, 88, 1061-1071.
- ROMERO, M. (1992). Ethnographic Evaluation of the Behavioral Causes of Census Undercount of Undocumented Immigrants and Salvadorans in the Mission District of San Francisco, California. Ethnographic Evaluation of the 1990 Decennial Census, Report #18. Prépare sous le Joint Statistical Agreement 89-41 avec San Francisco State University Foundation. U.S. Bureau of the Census, Washington, D.C.
- SHAPIRO, G., DIFFENDAL, G., et CANTOR, D. (1993). Survey undercoverage: major causes and new estimates of magnitude. *Proceedings of the 1993 Annual Research Conference*, U.S. Bureau of the Census, 638-663.
- STACK, C.B. (1974). *All Our Kin: Strategies for Survival in the Black Community*. New York: Harper and Row.
- SWEET, E.M. (1994). Roster research results from the living situation survey. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 415-433.
- U.S. DEPARTMENT OF LABOR (1994). Employment and Earnings. Bureau of Labor Statistics, janvier 1994; Washington, D.C.
- VALENTINE, C., et VALENTINE, B. (1971). Missing Men: A Comparative Methodology Study of Underenumeration and Related Problems. Rapport au U.S. Bureau of the Census, le 3 mai 1971.

sont censées être parties pour un voyage d'affaires. Ces personnes étaient considérées «de manière intuitive» comme faisant partie du ménage dont elles étaient absentes (Gerber, Wellens et Keeley 1996).

7. CONCLUSIONS

La mobilité temporaire, telle qu'elle est définie dans notre étude, comporte des déplacements de longue et de courte durée, fréquents et peu fréquents, prédéterminés et non prédéterminés, à partir d'un seul lieu de résidence, et ces déplacements comportent souvent un retour au lieu de résidence en question. Depuis longtemps, on pense que cette mobilité contribue à l'erreur de couverture qui entache les recensements et les enquêtes, par la confusion qu'elle crée lorsqu'il s'agit de déterminer qui vit ou habite au juste dans un lieu de résidence (ménage) donné.

Notre échantillon de personnes ayant des liens plus qu'occasionnels avec un ménage a montré un degré assez important de mobilité temporaire durant une période de deux à trois mois. Nous avons relevé des différences démographiques intéressantes dans le degré de mobilité, ainsi que dans les types de mobilité qui ont été déclarés. Nous avons constaté notamment que les membres du groupe «difficile à recenser» (hommes noirs/hispaniques âgés de 18 à 29 ans) étaient concentrés dans les catégories «boomerangs» et «volants», ce qui laisse penser qu'il existe une mobilité temporaire de type répétitif. Nous pensons que ces catégories comprennent des personnes qui ont des liens solides avec de multiples ménages (par ex., un fils adulte qui partage son temps entre la résidence d'un parent et celle d'une petite amie, ou une jeune mère qui séjourne périodiquement au sein de ménages appartenant à un réseau de parents proches afin de recevoir de l'aide pour prendre soins des enfants).

Outre l'inclusion des personnes qui ont tendance à échapper au dénombrement lors de recensements et d'enquêtes, d'autres considérations indiquent que les catégories des «boomerangs» et des «volants» présentent un intérêt particulier. Premièrement, comparativement à d'autres catégories de mobilité, les personnes faisant partie de ces groupes passent davantage de temps loin du ménage au sein duquel elles ont été «trouvées», et, deuxièmement, elles sont plus susceptibles d'avoir comme lieu de résidence habituel un endroit autre que le ménage en question. Il est difficile d'estimer dans quelle mesure ce type de mobilité contribue au sous-dénombrement. Cependant, il est très utile de noter que la moitié de la population difficile à recenser tombe soit dans la catégorie «boomerangs», soit dans la catégorie «volants». Cela semble être plus qu'une coïncidence qu'un segment aussi important de cette population fasse partie de l'un ou l'autre de ces deux groupes de mobilité, que l'on pourrait qualifier d'«ambigus» en ce qui a trait au lieu de résidence.

L'analyse log-linéaire laisse penser qu'il n'y a pas une relation claire et simple entre la mobilité temporaire et l'attribution du lieu de résidence habituel. Ainsi, nous n'avons pas constaté que plus le degré de mobilité temporaire est élevé, moins il y a de chances d'être classé comme résident habituel. La relation semble plutôt dépendre du type de déplacement. Par exemple, le voyageur de commerce ou le chauffeur de camion qui déclare la plus grande diversité d'endroits visités et le plus grand nombre de visites peut néanmoins présenter moins d'ambiguïté en ce qui a trait au lieu de résidence qu'une personne qui visite un seul endroit mais à plusieurs reprises. Or, c'est justement ce qu'on a constaté dans le cas de la catégorie «pas de répétition», où le temps passé loin du ménage n'avait pratiquement pas de rapport avec l'attribution de la résidence habituelle.

Notre examen de la mobilité temporaire représente une nouvelle orientation de recherche pour l'étude de l'erreur de couverture au sein des ménages lors de recensements et d'enquêtes. Nous formulons deux recommandations pour l'amélioration de la couverture des recensements et des enquêtes. Premièrement, les organismes d'enquête devraient étudier la possibilité de mesurer directement le rapport entre la mobilité temporaire et les cas de sous-dénombrement liés aux recensements et aux enquêtes. Cela pourrait se faire en ajoutant des questions portant sur la mobilité dans les interviews postcensitaires relatives à la couverture qui sont utilisées pour estimer le nombre de personnes non recensées ou dénombrées par erreur. Si la corrélation entre l'erreur de couverture et la mobilité est importante, les méthodes d'enquête pourraient alors être adaptées afin d'essayer de réduire cette erreur. Par exemple, de nouvelles questions relatives à la liste de départ pourraient être ajoutées aux formulaires de recensement et aux interviews de suivi portant sur la non-réponse, le but étant de trouver davantage de «boomerangs» et de «volants». Des mesures de la mobilité temporaire pourraient également se révéler de puissantes variables prédictives lors de la modélisation statistique du sous-dénombrement. Bien que nous ne soyons qu'au début des recherches, la mobilité temporaire semble un aspect prometteur en vue de mieux comprendre l'erreur de couverture survenant au sein des ménages.

REMERCIEMENTS

Les auteurs remercient Elizabeth A. Martin, Theresa J. DeMaio, Robert E. Fay et trois réviseurs anonymes pour les judicieux commentaires formulés au sujet de versions antérieures de la présente communication. Cet article fait état des résultats de recherche et d'analyse effectuées par le Census Bureau. Il fait état d'une revue plus limitée que ce que rapportent les publications officielles du Census Bureau. Cet article a pour but d'informer les équipes de recherche intéressées et de susciter la discussion.

Table 5
Test de qualité de l'ajustement et estimations de paramètres pour les modèles log-linéaires relatifs à l'incidence du sexe (S), de la mobilité temporelle (M) et du temps passé ailleurs (T) sur la détermination du statut de résident habituel (U)

A. Test de qualité de l'ajustement				
Modèle	d.l.	écart-type	(U) statut de résident habituel	
			chi carré +	p
1. H, SMT	15	4,79		0,00
2. HS, HM, HT, SMT	10	1,06		0,12
3. HTM, HSM, SMT	4	0,78		0,16
B. Estimations de paramètres, modèle 3				
(M) MOBILITÉ:	bêta	écart-type	valeur de référence	
«Une fois»	1,08	0,40		2,71*
«Boomerangs»	-1,54	0,39		-3,97*
«Pas de répétition»	0,83	0,58		1,43
«Volants»	-0,38	0,47		-0,80
(S) SEXE:				
(Hommes)	0,39	0,27		1,44
(T) TEMPS AILLEURS:				
(> ½ de la période de référence)	-1,78	0,27		-6,52*
INTERACTION (H)*(S)*(M) (hommes)				
«Une fois»	-0,64	0,43		-1,48
«Boomerangs»	0,69	0,58		1,18
«Pas de répétition»	0,85	0,62		1,37
«Volants»	-0,90	0,42		-2,14*
INTERACTION (H)*(M)*(T) (> ½ de la période de référence)				
«Une fois»	-0,72	0,48		-1,50
«Boomerangs»	-1,20	0,54		-2,26*
«Pas de répétition»	1,57	0,74		2,12*
«Volants»	0,36	0,41		0,88
* Test du chi carré «jackknife» pour l'ajustement global.				
* Significatif au niveau 0,5 %.				

étaient moins susceptibles d'être classées comme résidents habituels que les personnes de sexe féminin ($b = -0,90$). Une analyse approfondie a révélé peu d'indices permettant d'expliquer cette constatation. Les «volants» de sexe masculin et de sexe féminin présentaient peu de différences en ce qui a trait aux types d'endroits visités, aux raisons des visites et à la relation à l'égard du chef de ménage des endroits visités (parent ou non parent). L'interaction reflète peut-être des différences ayant trait à d'autres liens sociaux, comme la présence d'enfants ou de biens personnels, ou l'apport de ressources.

La base du tableau 5 montre que l'interaction entre le lieu de résidence habituel, la mobilité et le temps passé loin du ménage est plutôt complexe. Le temps passé loin du ménage sondé semble avoir une incidence sur la détermination du statut de résident habituel, mais seulement dans le cas de certains types de mobilité. Le coefficient d'interaction est important et négatif dans le cas des «boomerangs» ($b = -1,20$). Ainsi, les probabilités d'être classé comme résident habituel sont encore plus faibles dans le cas des «boomerangs» qui se sont absentes ou ménages durant la moitié de la période de référence ou davantage, comparativement aux membres d'autres groupes

Cependant, dans le cas de la catégorie «pas de répétition», le coefficient est important et positif, ce qui neutralise à toutes fins pratiques l'effet principal négatif du temps passé loin du ménage sondé ($1,57 + -1,78 = -0,21$). Dans le cas de ce groupe, le temps passé loin du ménage ne semble pas être lié à l'attribution du statut de résident habituel. Il semblerait que des facteurs autres que le temps passé ailleurs pourraient être plus importants dans le processus cognitif de détermination du lieu dans lequel ces personnes «résident». Une hypothèse à cet égard est que les personnes faisant partie de la catégorie «pas de répétition» sont des gens qui doivent voyager pour leur travail et qui, manifestement, ont un lieu de résidence stable malgré leurs déplacements fréquents et les longues périodes d'absence du ménage. Cette hypothèse était les constatations d'une brève étude qui a révélé que les répondants n'avaient pas besoin d'une règle énoncée pour être en mesure d'identifier correctement le lieu de résidence habituel de personnes qui

ou davantage. La race et l'âge ont été exclus étant donné qu'aucune de ces deux variables n'améliorait la qualité de l'ajustement des données.

Le modèle de base (H, SMT) ne donnait pas un bon ajustement des données; c'est pourquoi nous avons rejeté l'hypothèse nulle, c'est-à-dire l'hypothèse selon laquelle l'attribution du statut de résident habituel est indépendante du type de mobilité, du sexe et du temps passé loin du ménage (X^2 «jackknife») pour la qualité globale de l'ajustement = 4,79, $d.l. = 15$, $p = 0,00$, voir tableau 5). Nous avons par la suite adapté un modèle (2) d'effets principaux qui comprenait les effets additifs de S, M et T sur H (HS, HM, HT, SMT). Ce modèle a donné un bon ajustement (X^2 «jackknife») pour la qualité globale de l'ajustement = 1,06, $d.l. = 10$, $p = 0,12$). Enfin, nous avons également adapté un modèle (3) comprenant deux termes d'interaction (HTM, HSM, SMT). Ce modèle présuppose des effets interactifs de T*M et de S*M sur H. Une comparaison entre le modèle d'effets principaux et le modèle d'effets interactifs laissait penser que plusieurs interactions étaient significatives et devaient être retenues (X^2 jackknife) pour la comparaison = 1,99, $d.l. = 6$, $p = 0,02$). Le tableau 5 indique la qualité globale des tests d'ajustement, ainsi que les estimations de paramètres tirées du modèle d'interaction le mieux ajusté (HTM, HSM, SMT - X^2 «jackknife») pour la qualité globale de l'ajustement = 0,78, $d.l. = 4$, $p = 0,16$). Les estimations de paramètres tirées du tableau 5 montrent que la mobilité temporaire a un effet principal significatif sur l'attribution du statut de résident habituel dans le modèle 3, qui tient compte du sexe, du temps passé loin du ménage et de plusieurs interactions. Deux des catégories de mobilité présentaient des coefficients bêta significatifs, bien que les directions étaient opposées. Les personnes de la catégorie «une fois» étaient nettement plus susceptibles d'être classées comme résidents habituels ($b = +1,08$). À l'inverse, les «boomerangs» présentaient une estimation de paramètre négative ($b = -1,54$), ce qui signifie que les probabilités d'être classé comme résident habituel étaient réduites de manière importante dans le cas de ce groupe.

Le temps passé ailleurs qu'au sein du ménage sondé avait de loin l'effet le plus important sur la détermination du statut de résident habituel, avec une forte association négative ($b = -1,78$). Cela veut dire que dans le cas de notre population qui présente une mobilité temporaire, les personnes absentes du ménage durant la moitié de la période de référence ou davantage étaient nettement moins susceptibles d'être considérées comme des résidents habituels que celles qui étaient absentes moins de la moitié de la période en question. La variable «sexe» n'avait pas eu un effet principal important, mais contribuait à une interaction significative. L'interaction apparaît dans le groupe des «volants», où les personnes de sexe masculin

ménage, du lieu de résidence habituel. Cette analyse va au-delà des constatations descriptives de la typologie, afin de déterminer si la mobilité a une incidence sur la perception, par le chef du ménage, de la notion de résidence. L'attribution du lieu de résidence habituel a été utilisée comme variable dépendante dans une série de modèles. La variable dépendante était composée de deux catégories: 1) résident habituel au sein du ménage sondé; 2) non résident habituel au sein du ménage sondé. Les personnes considérées comme des résidents habituels tant au sein du ménage sondé qu'à un autre endroit ont été regroupées dans la première catégorie. Les variables prédictives étaient les suivantes: âge, sexe, race, temps passé ailleurs et types de mobilité. Les modèles finals présentés au tableau 5, qui comprennent tous les termes relatifs à l'interaction des variables indépendantes, équivalent à des modèles logit pour le lieu de résidence habituel.

Le premier modèle testait la mobilité comme mesure dichotomique: les personnes sans mobilité (les «non-mobiles») et celles qui avaient passé au moins une nuit loin du ménage sondé (les catégories «une fois», «pas de répétition», «boomerangs» et «volants» réunies). Ce modèle devait établir d'abord si la mobilité temporaire était un prédicteur important du statut de résidence, tous types de mobilité confondus. Cette «première coupe» était nécessaire parce qu'environ 50 % de l'échantillon se trouvait dans la catégorie «non-mobiles» et parce que cette catégorie était fortement désaxée vers la catégorie «résident habituel» de la variable dépendante. Par conséquent, les modèles avec lesquels on essayait d'inclure toutes les cinq catégories de la typologie de mobilité étaient mal spécifiés en raison du grand nombre de cellules garnies de zéros.

Le modèle comportant la mesure dichotomique de la mobilité et la variable «sexe» a permis d'obtenir des données relativement bien «ajustées» (X^2 «jackknife») pour la qualité globale de l'ajustement = 0,28, $d.l. = 2$, $p = 0,27$). Ni la race ni l'âge n'ont amélioré l'ajustement. Les estimations de paramètres ont montré que les personnes faisant partie de la catégorie «non-mobiles» étaient plus susceptibles d'être classées comme résidents habituels que les personnes ayant une certaine mobilité (données non indiquées).

Ayant déterminé que la mobilité était liée de manière significative au choix du statut de résident, nous avons cherché à savoir si le type de mobilité temporaire était un prédicteur. Tout d'abord, nous avons testé un modèle à données de base indépendantes, afin de prédire la résidence habituelle (H). Les variables prédictives étaient les suivantes: mobilité (M), sexe (S) et temps passé loin du ménage (T). La variable de la mobilité était composée des quatre types de mobilité («une fois», «pas de répétition», «boomerangs» et «volants»). Le temps passé loin du ménage a été séparé en deux catégories: moins de la moitié de la période de référence, et la moitié de la période de référence

5. LIEU DE RÉSIDENCE HABITUEL ET MOBILITÉ

Inversement, nous pouvons présupposer, par définition, que les «boomerangs» qui étaient absents plus de la moitié de la période de référence ont passé la majeure partie de leur temps au seul autre endroit qu'il ont déclaré avoir visité. En supposant que le temps joue un rôle dans la définition d'un sens d'appartenance à un ménage, on peut présupposer que les «boomerangs» ont une meilleure chance d'être recensés car ils passent la plupart de leur temps à l'autre endroit.

Nous avons essayé de déterminer si la mobilité temporelle a une incidence sur la désignation, par le répondant du ménage, des «résidents habituels». Sur le formulaire de Recensement américain de 1990, on demandait aux répondants de dresser la liste des personnes à l'endroit où celles-ci habitaient ou dormaient la plupart du temps, tandis que la LSS demandait aux répondants s'ils considéraient le ménage sondé comme le «lieu de résidence habituel, c'est-à-dire l'endroit où [vous/NOM] vivez(vit) et dormez(dort) la plupart du temps». Les répondants étaient également invités à répondre à la question suivante: «Est-ce que [vous/NOM] avez(a) une résidence habituelle ailleurs?». Bien que cette méthode ne soit pas une réplique parfaite d'une liste de questionnaire de recensement, elle permet de déterminer de manière approximative quelles personnes parmi celles mentionnées sur la liste de la LSS auraient pu être normalement incluses dans un questionnaire de recensement ou d'enquête courante, ou exclues de celui-ci, par le chef du ménage.

Le tableau 4 présente une classification croisée de l'attribution de la résidence habituelle d'après le statut de mobilité. Une combinaison des questions habituelles relatives au lieu de résidence a donné quatre possibilités de classement: résidence habituelle seulement au ménage sondé; résidence habituelle à un quelconque endroit autre que le ménage sondé; résidence habituelle à un autre endroit; résidence habituelle à cet endroit. En présupposant que les réponses d'«autre endroit» équivaient à être exclu du formulaire de recensement, nous constatons que dans l'ensemble, seulement 4 % des personnes qui entretenaient des liens «plus qu'occasionnels» à l'égard du ménage sondé pouvaient avoir été ainsi exclues. Globalement, la répartition des classifications de résident habituel diffèrait de manière significative selon le type de mobilité.

Comme on peut s'y attendre, presque toutes les personnes qui ont passé chaque nuit de la période de référence au sein du ménage ont été considérées comme des résidents habituels à cet endroit (chiffre arrondi à 100 %). L'écart le plus évident parmi les diverses catégories touche les «boomerangs» et les «volants». Entre 20 et 25 % des personnes faisant partie de ces deux catégories ont été désignées par les répondants des ménages comme étant des résidents habituels d'un endroit autre que le ménage sondé.

Cette situation est très différente tant de celle du groupe des «une fois» que de celle de la catégorie «pas de répétition», où seulement 2 % et 5 % de personnes respectivement ont été considérées comme des résidents habituels d'un autre endroit. Ces résultats laissent penser que les deux derniers groupes reflètent un type de mobilité associé aux loisirs ou aux affaires, mais dans le cas de personnes ayant des liens solides à l'égard du ménage, tandis que les groupes des «boomerangs» et des «volants» sont plus susceptibles de comprendre des personnes qui ont des liens moins forts avec le ménage. Pour cette raison, et parce qu'un pourcentage important de la population difficile à dénombrer se trouvait dans ces deux catégories, on peut dire que les «boomerangs» et les «volants» présentent les aspects les plus intéressants en ce qui a trait à la couverture et soulèvent plusieurs questions. Par exemple, est-ce que ces personnes sont recensées à un seul endroit, à tous les endroits ou à aucun endroit? Où devrait-on les recenser?

Tableau 4
Où le répondant du ménage considère-t-il que la personne visée est un «résident habituel»? (Pourcentage pondéré et écarts-types)

Résident habituel: où?	Non-mobiles	Une fois	Boomerangs-rangs	Pas de répétition	Volants	Total
Ménage sondé	100 %	97 %	71 %	95 %	70 %	95 %
seulement	(0,2)	(2,0)	(12,1)	(4,2)	(10,0)	(1,7)
Autre endroit	0 %	2 %	25 %	5 %	20 %	4 %
Aux deux endroits	0 %	1 %	4 %	0 %	10 %	1 %
endroits	(-)	(0,4)	(2,1)	(-)	(7,3)	(0,5)
Nombre non pondéré	716	314	186	101	134	1451

Test du chi carré («Jackknife») = 2,79, p < 0,05, d.l. = 8

Le fait qu'un pourcentage relativement important des «boomerangs» et des «volants» soient considérés comme des résidents à un quelconque endroit autre que le ménage sondé laisse voir la possibilité d'un sous-dénombrement. D'autre part, le fait que 10 % des «volants» soient désignés comme des résidents habituels tant au sein du ménage sondé qu'à un autre endroit laisse penser qu'il y a un risque de surdénombrement. Dans ces situations d'incertitude quant au lieu de résidence (résidents habituels ailleurs ou aux deux endroits), le nombre pondéré de «boomerangs» et de «volants» représente environ 4 % de la population totale. De ce point de vue plus général, il semble qu'un segment non négligeable de la population peut faire l'objet d'une erreur de couverture.

6. MODÉLISATION DU LIEU DE RÉSIDENCE HABITUEL ET DE LA MOBILITÉ

Dans la présente section, nous effectuons une modélisation statistique de la détermination, par le répondant du

soit le lieu de résidence pour le reste de l'année. Gerber (1994) a constaté que les répondants utilisent également le critère du temps à des degrés différents pour dresser la liste des personnes faisant partie du ménage; elle a notamment constaté que dans certaines situations, il n'y a pas de rapport clair entre le fait de figurer sur la liste du question-naire et le temps passé à un endroit donné. Ce sont plutôt des facteurs comme l'appartenance à un ménage et les liens à l'égard de celui-ci qui semblent avoir une plus grande incidence sur le processus de prise de décisions. Cependant, il est sensé de penser, de manière intuitive, que le temps passé ailleurs a une certaine incidence sur la décision du chef de ménage lorsqu'il s'agit de déterminer qui doit être inclus dans la liste du questionnaire. Afin de déterminer quelle est l'incidence du temps passé ailleurs sur nos catégories de mobilité, nous avons divisé le nombre total de nuits passées loin du ménage sondé au cours de la période de référence par le total des nuits de cette période; le résultat est exprimé sous forme de pourcentage. Le tableau 3 présente cette mesure de temps sous une forme indiquant une absence supérieure ou inférieure à la moitié de la période de référence.

Tableau 3

Temps passé loin du ménage sondé durant la période de référence (Pourcentage pondéré et écarts-types)				
Loi 50 % du temps ou Une fois Boomerangs Pas de répétition				
Total				
Volants				
Non				
94 %	(4,4)	73 %	(11,5)	98 %
6 %	(4,4)	27 %	(11,5)	2 %
12 %	(3,6)	37 %	(10,3)	63 %
12 %	(3,6)	37 %	(10,3)	88 %
Oui				
314	186	101	134	735
Test du chi carré («Jackknife») = 1,71, p. < 0,05, d.l. = 3				

Les groupes des «boomerangs» et des «volants» étaient plus susceptibles que les autres groupes de passer la moitié ou plus de la période de référence à un endroit autre que le ménage sondé. Cette constatation vient étayer l'hypothèse selon laquelle les visites répétées qui caractérisent ces deux groupes sont associées à une augmentation du temps total passé loin du ménage sondé. Elle laisse penser également que ces groupes présentent une plus grande ambiguïté en ce qui a trait au lieu de résidence, en particulier le groupe des «volants». Étant donné que les membres de ce groupe déclarent avoir visité au moins deux endroits, outre le ménage sondé, il n'est pas clair si les personnes qui s'absentent plus de la moitié du temps passent la plupart du temps en un endroit donné. Si ces personnes passent à peu près le même temps à chacun des endroits, il est facile d'imaginer que les «volants» ne soient mentionnés sur aucune liste, à aucun endroit, ou sur plus d'une liste.

l'échantillon LSS répondait au critère relatif à la difficulté de dénombrer, mais l'analyse de la mobilité de ce sous-groupe révèle des caractéristiques très différentes des celles du sous-groupe «non difficile à dénombrer» (non DD).

Premièrement, le groupe DD semble être plus mobile (plus de 60 % des répondants ont indiqué avoir passé au moins une nuit à un endroit autre que le ménage sondé, comparativement à un taux de 50 % dans le cas du groupe non DD). Deuxièmement, la répartition des catégories de mobilité diffère de manière significative si l'on tient compte du critère relatif à la difficulté de dénombrer. La plupart des répondants du groupe non DD qui sont mobiles sont concentrés dans la catégorie «une fois», tandis que les répondants mobiles du groupe DD sont plutôt concentrés dans les catégories à déplacements répétés («boomerangs» et «volants», avec des valeurs de 21 % et 29 % respectivement).

Nous avons aussi étudié la répartition de la mobilité temporaire selon la race (Blancs, Noirs, Hispaniques et autre) et l'âge (0-17, 18-29, 30-49, 50 ans et plus). Dans l'ensemble, la mobilité temporaire ne varait pas de manière significative, ni par race, ni par âge. On a constaté une concentration relativement importante d'Hispaniques dans la catégorie «pas de répétition» (19 %) et de Noirs, dans la catégorie «volants» (9 %). Les Noirs étaient plus nombreux que les Blancs dans la catégorie «non mobiles» (66 % comparativement à 52 %), et ce malgré le fait que les Noirs présentaient un taux de mobilité géographique plus élevé que les Blancs. Enfin, les jeunes adultes âgés de 18 à 29 ans étaient plus mobiles que les répondants appartenant à d'autres groupes d'âge (près de 70 % des répondants de ce groupe d'âge avaient passé au moins une nuit loin du ménage sondé) et un pourcentage disproportionné de répondants de ce même groupe faisaient partie de la catégorie «volants» (14 %). L'absence de signification statistique dans le cas de certaines de ces tendances pourrait être un artefact dû à la taille de l'échantillon. Ou alors, la mobilité temporaire peut être suffisamment différente de la mobilité géographique pour que les caractéristiques ne soient pas les mêmes que celles observées dans le cas des gens qui déménagent souvent.

Une autre variable importante que nous avons présumée être corrélée avec le modèle de mobilité temporaire est le temps passé loin du ménage sondé pour des visites. Les règles du U.S. Census Bureau relatives au lieu de résidence habituel varient selon un critère de temps. Ainsi, les personnes qui travaillent dans une autre ville durant la semaine et qui retournent à la maison au cours du week-end doivent être recensées à l'endroit où elle «vit» et dortiment la plupart du temps (dans ce cas, à l'endroit où elles vivent en semaine). Toutefois, un(e) enfant qui vit dans un pensionnat doit être recensé(e) au lieu de résidence de ses parents, même s'il ou elle passe probablement la plupart du temps à l'école. De la même manière, une personne qui, le jour du recensement, vit dans un logement de groupe (par ex., une résidence d'étudiants universitaires ou une prison) est recensée à l'endroit en question, quel que

destinations multiples" en comptant comme un déplacement distinct chaque endroit visité durant un seul déplacement. Nous avons donc mis au point une seule mesure de la mobilité qui applique diverses combinaisons du nombre d'endroits et du nombre de déplacements. En tout, nous avons créé cinq catégories afin d'identifier divers types de déplacements, et de distinguer les personnes qui effectuent des visites répétées au même endroit. Notre première catégorie désigne les personnes qui ont passé toutes les nuits de la période de référence au sein du ménage sondé et qui n'ont pas de mobilité temporaire (les «non-mobiles»). La deuxième catégorie regroupe les personnes qui, d'après le calendrier, ont déclaré une seule visite à un seul endroit (les «une fois»). La catégorie des «boomerangs» regroupe les personnes qui ont fait des visites répétées à un seul endroit. La catégorie «pas de répétition» désigne les personnes qui se sont rendues à plus d'un endroit, mais jamais deux fois au même endroit. Enfin, les «volants» sont les personnes qui ont passé la nuit à différents endroits et qui sont retournées plusieurs fois à au moins un de ces endroits (voir tableau 1).

Tableau 1
Typologie de la mobilité temporaire

Nombre d'endroits visités	Nombre de visites			
	0	1	2	3
Non-mobiles				
1	Une fois	Boomerangs	Boomerangs	Boomerangs
2	Pas de répétition	Volants	Volants	
3	Pas de répétition	Volants		
4	Pas de répétition			

4. CARACTÉRISTIQUES DES TYPES DE MOBILITÉ

Le tableau 2 montre la fréquence pondérée des divers types de mobilité. Un peu plus de la moitié des personnes qui ont rempli le questionnaire individuel n'ont déclaré aucune mobilité en dehors du ménage sondé pour la période de référence. Le plus grand nombre de personnes mobiles se trouve dans la catégorie «une fois»; ces personnes ont déclaré avoir fait un seul déplacement à partir du ménage sondé à destination d'un autre endroit (26 % au total). La catégorie des «boomerangs» (11 %) réunit les personnes ayant effectué deux visites ou plus à un seul endroit, tandis que 7 % des répondants se sont situés dans la catégorie «pas de répétition», qui est caractérisée par une mobilité non répétitive mais fréquente. Quant aux «volants», ils constituent le plus petit groupe (4 %).

Le tableau 2 présente également des données démographiques choisies relatives aux cinq catégories de mobilité, y compris la répartition selon le sexe, qui indique une plus

grande tendance à la mobilité chez les hommes que chez les femmes. Environ 60 % des hommes ont déclaré au moins une visite en dehors du ménage sondé, ce qui est un pourcentage nettement plus élevé que dans le cas des femmes (environ 33 %). Cette différence entre les sexes en ce qui a trait à la mobilité temporaire est beaucoup plus marquée que dans le cas de la mobilité géographique, où l'écart entre les hommes et les femmes n'est que d'environ 1 % (17 % des hommes ont déménagé entre 1993 et 1994, tandis que 16 % des femmes ont fait de même durant cette période; voir Hansen 1994). Cela laisse penser que la mobilité temporaire est un phénomène plus courant que la mobilité géographique, et que les caractéristiques démographiques associées à ces phénomènes sont également différentes. Les déplacements liés au service militaire pourraient expliquer la différence entre les sexes en ce qui a trait à la mobilité temporaire, tout comme les voyages d'affaires, les hommes présentant des rapports service actif/population et emploi/population plus élevés que les femmes (U.S. Department of Labor 1994).

Tableau 2
Typologie de la mobilité temporaire par sexe et par difficulté de dénombrement (difficile à dénombrer – DD)*

TYPE DE MOBILITÉ	Pourcentage total pondéré (écarts-types entre par.)	SEXE		DIFFICULTÉ DE DÉNOMBREMENT
		MASC.	FÉM.	
Non-mobiles	52 % (14,0)	40 % (13,7)	67 % (13,6)	38 % (7,8)
Une fois	26 % (10,4)	35 % (13,9)	16 % (7,0)	6 % (2,9)
Boomerangs	11 % (4,0)	15 % (5,7)	6 % (2,9)	21 % (9,1)
Pas de répétition	7 % (2,9)	6 % (2,4)	8 % (4,3)	6 % (5,4)
Volants	4 % (1,0)	4 % (1,3)	3 % (0,9)	29 % (9,9)
Nbre non pondéré	1 451	653	798	76

χ^2 pour la distribution, excluant la catégorie «non-mobiles» = 2,14, $p > 0,05$, $d.l. = 4$

* Test du chi carré (χ^2) = 2,03, $p > 0,05$

* Le groupe des «difficiles à dénombrer» comprend les hommes noirs et hispaniques âgés de 18 à 29 ans. ** Voir Fay 1985 pour l'application du test du chi carré («jackknife») à des échantillons complexes.

La partie droite du tableau 2 réunit plusieurs caractéristiques démographiques afin de créer un sous-groupe dont on sait qu'il présentait un taux de sous-dénombrement élevé lors de recensements précédents. Ce sous-groupe est composé d'hommes noirs ou hispaniques âgés de 18 à 29 ans. Ce sous-groupe est souvent qualifié de population «difficile à dénombrer» (DD). Seul un faible pourcentage de

influencent sur le jugement du répondant d'un ménage lorsqu'il s'agit de déterminer qui est un résident habituel du ménage en question. Le temps passé loin du ménage, le nombre de visites et le nombre de destinations sont une mesure indirecte de la force des liens précités.

Notre typologie de la mobilité temporaire a été créée à l'aide de deux aspects relatifs aux déplacements comportant des nuits passées à l'extérieur du ménage sondé. Le premier aspect a trait à la diversité des endroits qu'une personne a visités au cours de la période de référence. Cet aspect donne une idée du nombre d'endroits autres que le ménage sondé avec lesquels une personne peut avoir des liens. Le deuxième aspect a trait à la fréquence des déplacements effectués à partir du ménage sondé, c'est-à-dire qu'on compte le nombre de fois qu'une personne a quitté le ménage pour une nuit ou davantage.

L'utilisation de ces facteurs comme mesure de la solidité des liens existant à l'égard d'un ménage est renforcée par des descriptions ethnographiques de modes de vie à forte mobilité. Le modèle de déplacements représenté dans notre typologie reflète un grand nombre de processus sociaux différents, comme les liens lâches existant à l'égard de ménages composés de parents proches (Stack 1974; Dressler Hoepfner et Pitts 1985), les modèles d'immigration (Wingard 1992), et l'adaptation à la pauvreté (Hainner 1987; Valentine et Valentine 1971).

La LSS comportait plusieurs questions exploratoires ouvertes dont le but était d'étudier la perception qu'avaient les répondants des raisons de leur mobilité. Par ces questions on voulait connaître la raison de certains déplacements (départs et retours). Nous avions espéré que ces questions allaient nous permettre d'évaluer de façon plus directe les structures sociales sous-jacentes qui sont à l'origine de la mobilité temporaire. Malheureusement, les réponses à ces questions étaient difficiles à décoder sans faire de présuppositions injustifiées; cette difficulté était due principalement à la façon dont avaient été formulées les réponses. C'est pourquoi nous n'avons pas tenu compte de ces réponses dans l'établissement de la typologie.

Chaque «déplacement» a été défini comme un séjour effectué loin du ménage sondé durant au moins une nuit. Ainsi, par exemple, si une personne était partie pour passer trois jours chez une amie, puis une nuit chez un parent, avant de revenir au ménage sondé, on attribuait à la personne en question deux endroits et deux visites (une visite pour chaque endroit), tandis que si une personne avait passé une nuit chez un ami avant de revenir au ménage et retourner deux semaines plus tard à la maison de ce même ami pour une deuxième visite, on attribuait à la personne en question un endroit et deux visites (deux visites consécutives). Le premier exemple montre un biais potentiel de cette méthode: celui qui consiste à compter chaque endroit distinct visité durant un départ prolongé du ménage sondé comme un déplacement indépendant (comme des vacances avec des destinations multiples). D'autre part, cette méthode saisisait également les déplacements des «voyageurs à

personnes représentant des gens qui avaient des liens plus que fortuits avec les ménages sondés et qui sont inclus dans les analyses présentées ci-après ($N = 1\ 451$).

La partie individuelle du questionnaire a donné un taux de réponse de 85,3 %. La majorité des interviews individuels ont été menées en personne (96 %), et la plupart des adultes (89 %) ont répondu pour eux-mêmes tandis que toutes les interviews d'enfants ont été réalisées auprès des adultes des personnes qui connaissaient bien les enfants. Étant donné que les chefs de ménage étaient invités à répondre à des questions de base sur le mode de vie et à des questions démographiques pour *toutes* les personnes figurant sur la liste établie au départ, nous disposons de certains moyens pour étudier les caractéristiques des quelque 15 % de personnes choisies pour remplir le questionnaire mais qui n'ont pas répondu. Nous n'avons pas constaté de différences significatives d'après l'âge ou le sexe entre les non-répondants et les répondants; en revanche nous avons relevé qu'un pourcentage disproportionné de non-répondants était de race noire. Nous avons constaté également que les non-répondants étaient plus susceptibles que les répondants d'avoir passé plus d'une semaine loin du ménage sondé. Ces constatations permettent de déterminer en partie le degré de représentativité de notre échantillon individuel, tant du point de vue démographique qu'en ce qui a trait à la mobilité temporaire. Étant donné que les non-répondants étaient plus susceptibles de s'absenter du ménage que les répondants, nous soupçonnons que le biais potentiel de «sélection» pourrait avoir causé une sous-estimation dans le cas de nos mesures de la mobilité.

Nous avons appliqué des poids individuels et de ménage afin de tenir compte du surechantillonnage, du critère de sélection pour l'enquête individuelle et de la non-réponse (voir Lynch, Witt, Branson et Ardini 1993). Toutes les analyses ont été effectuées à l'aide du programme de tableaux de contingence pour plans d'échantillonnage complexes (Contingency Table Analysis for Complex Sample Designs – CPLX), qui est un programme informatique d'estimation de la variance conçu pour tenir compte des effets du plan d'échantillonnage complexe de la LSS (voir Fay 1989b, 1985).

3.1 Typologie de la mobilité temporaire

La typologie que nous présentons ici a été établie de manière empirique, c'est-à-dire que le regroupement de visites et de destinations a été obtenu par voie analytique et non théorique. Par conséquent, les catégories que nous identifions ne représentent pas des groupes de personnes ayant des caractéristiques identiques ou se trouvant dans des situations identiques. Cette typologie doit plutôt être considérée comme un essai visant à représenter la réalité complexe sous-jacente à des modes de vie comportant de la mobilité. Notre hypothèse est qu'une telle mobilité a une incidence sur la force des liens sociaux existant entre un individu et un ménage en particulier, et que ces liens

rées comme des visiteurs occasionnels. (Des 712 visiteurs occasionnels, 77 % avaient un lien de parenté avec le répondant du ménage sondé, 93 % étaient non Hispaniques, 84 % étaient Blancs et 58 % étaient de sexe féminin). Pour plusieurs raisons, ces personnes n'étaient pas «admissibles» pour le reste du questionnaire. Premièrement, nous avons jugé que les visiteurs occasionnels ne correspondaient pas à la définition d'un résident habituel utilisé par le Census Bureau, et, deuxièmement, l'exclusion de ce groupe de la majeure partie du questionnaire a permis de réduire de manière importante le temps et les ressources nécessaires à la réalisation de l'enquête.

Après le suivi destiné à obtenir une interview dans les cas où celle-ci avait été refusée ou n'avait pas eu lieu, le taux de réponse final pour la partie de l'interview menée au niveau du ménage a été de 79,5 %. (Un suivi afin de convaincre à la participation à l'enquête, a été effectué par lettre, par appel direct du superviseur de la collecte de données, par des visites répétées et par la distribution d'interviews. Les répondants ont été contactés en moyenne 1,9 fois, les non-répondants 5,9 fois). Compte tenu de la population sondée, ce résultat était considéré comme acceptable. Cependant, étant donné que nous soupçonnons que la non-réponse est liée de manière importante à des questions relatives à la couverture, comme la mobilité, il est probable que ce taux de non-réponse ait une certaine incidence sur nos estimations. Cet aspect est traité plus en détail dans la description du questionnaire qui est donnée ci-après.

La partie suivante de l'enquête comportait un questionnaire individuel à remplir soi-même et contenant des questions sur la mobilité temporaire ainsi que des données démographiques fournies par les répondants, auxquels on demandait s'ils avaient passé une nuit ailleurs que dans le lieu de résidence sondé, durant la période de référence. Dans l'affirmative, les intervieweurs utilisaient un calendrier pour noter les endroits et les dates des séjours effectués hors de la résidence habituelle. Les intervieweurs recueillaient également des renseignements sur chacun de ces endroits, ainsi que sur le type de lien existant entre le répondant et les lieux en question et sur la ou les raisons de ces séjours.

Chacun des chefs de ménage remplissait un questionnaire individuel pour lui-même ou elle-même. En outre, toutes les personnes figurant sur la liste établie au départ qui avaient séjourné ailleurs pendant huit jours ou plus au cours de la période de référence remplissaient, elles aussi, le questionnaire individuel. Toutes les personnes identifiées comme étudiante(s) universitaire, ainsi que les personnes qui n'avaient pas de résidence habituelle étaient également admissibles à une interview individuelle. Enfin, le questionnaire individuel a été soumis également à un simple échantillon aléatoire de 10 % des ménages visés par la LSS. Au sein de ces ménages, on a essayé d'effectuer une interview auprès de chaque personne figurant sur la liste, à l'exception des visiteurs occasionnels. Ce critère de sélection quelque peu complexe a donné un groupe de

patronnée par le U.S. Census Bureau entre mai et septembre 1993. L'échantillon avait été stratifié de manière à sur-échantillonner les régions à concentration élevée et moyenne de minorités (c'est-à-dire où plus de 80 % de la population est composée de Noirs ou d'Hispaniques, et où cette proportion se situe entre 40 et 80 %), ainsi que les régions habitées par des locataires (c'est-à-dire où les locataires constituent plus de 40 % de la population). Afin d'accroître l'efficacité du plan d'échantillonnage, le RTI avait utilisé des données relatives aux unités de logement qui avaient été recueillies antérieurement à partir d'un échantillon aléatoire à plusieurs degrés ayant servi dans le cadre de la National Household Survey on Drug Abuse (NHSDA) de 1992.

La première partie de l'interview de la LSS a été menée en personne avec le répondant du ménage qui était en mesure de donner le plus de renseignements, dans la plupart des cas, le chef du ménage (d'après la définition du U.S. Census Bureau, le chef du ménage est la personne qui possède ou loue la maison ou le logement). Les chefs de ménage fournissaient une liste de personnes et répondaient ensuite à des questions démographiques, pour eux(elles)-mêmes ainsi que pour toutes les autres personnes mentionnées sur la liste. Au moyen d'une série de questions approfondies, le questionnaire permettait de dresser la liste des résidents formant le «noyau» du ménage mais aussi d'inclure un grand nombre de personnes dont la présence au sein du ménage était moins durable. Les personnes qui avaient des liens plus fragiles étaient incluses dans le questionnaire en posant des questions qui cherchaient à savoir qui avait passé la nuit dans la maison ou le logement durant la période de référence, qui était considéré comme un membre du ménage même s'il ou elle habitait ailleurs, et qui considérait la résidence comme son adresse permanente ou comme lieu de réception d'envois postaux ou de messages téléphoniques (voir Sweet 1994). (La durée de la période de référence variait selon la date de l'interview. Les périodes de référence commençaient le premier jour du mois, 2 mois précédant le mois pendant lequel l'interview s'effectuait et se terminait le jour de l'interview. Par conséquent, les interviews effectuées à la fin du mois avaient une plus longue période de référence que les interviews effectuées au début du mois). Au total, on a effectué des interviews auprès de 999 ménages à l'échelle nationale. À l'aide de la méthode d'établissement d'une liste générale, on a pu dénombrer 3 549 personnes.

L'étape suivante de l'enquête consistait à écarter de la liste les personnes qui étaient considérées comme des «visiteurs occasionnels». Une personne était considérée comme un «visiteur occasionnel» si son lieu de résidence habituel était, de l'avis du chef du ménage, un endroit autre que l'unité de logement échantillonnée et si cette personne n'avait pas séjourné au sein du ménage durant plus d'une semaine au cours de la période de référence. Ce processus de sélection a permis d'identifier sur la liste les personnes qui n'avaient qu'un lien occasionnel avec le ménage sondé. Sur les 3 549 personnes dénombrées, 712 ont été considé-

2. RENSEIGNEMENTS D'ARRIÈRE-PLAN

Le déplacement d'un lieu géographique à une autre est généralement indiqué par un changement d'adresse, ainsi que par le déplacement de biens, etc. Ce type de mobilité est généralement désigné par le terme «mobilité géographique», il y a une forme de mobilité qui est moins bien définie: la mobilité temporaire. Par ce terme, nous désignons les déplacements temporaires et parfois prédéterminés qui peuvent être de longue ou de courte durée, fréquents ou peu fréquents, et qui peuvent comporter des nuitées. Ce type de mobilité a été décrit comme «une des caractéristiques clés des ménages irréguliers et complexes» («one of the key features of irregular and complex households», de la Puente 1993). On trouve un exemple de ce genre de ménage au sein des collectivités d'Haïtiens, où la structure type d'un ménage est composée d'un «noyau» relativement stable et d'une «périphérie fluide». Celle-ci est composée de nouveaux arrivants, apparentés ou non, qui séjournent au sein du ménage pour de courtes périodes, ainsi que de membres du ménage qui se rendent régulièrement en visite à Haïti pour des séjours de plusieurs semaines ou de plusieurs mois (Wingard 1992). La mobilité temporaire n'est pas un aspect que l'on trouve uniquement au sein de collectivités particulières. Il existe de nombreux exemples de ce type de mobilité dans la population en général, notamment une mobilité associée à des voyages d'affaires ou de vacances de longue durée, à la fréquentation d'universités, à des situations de garde et à des personnes qui sont présentes dans un ou plusieurs ménages durant une période donnée. Cette mobilité dans la périphérie fluide (mobilité temporaire) diffère de la mobilité géographique parce qu'elle consiste en des déplacements caractérisés par des départs d'un seule résidence qui sont généralement suivis d'un retour à celle-ci. Il est difficile pour les répondants d'un recensement ou d'une enquête de savoir quels membres de la périphérie fluide devraient être mentionnés comme faisant partie d'un ménage donné. En effet, les déplacements de ces personnes peuvent ne pas comporter un changement d'adresse permanente, ce qui peut rendre difficile la détermination du nombre de personnes qui vivent ou habitent à une adresse donnée.

Étant donné qu'il y a peu d'études qui portent sur la mobilité temporaire, les travaux qui ont pour objet la mobilité géographique et la structure des ménages constituent un bon point de départ pour la formulation de nos hypothèses concernant la mobilité temporaire. D'après les données recueillies lors de la Current Population Survey de mars 1994, les jeunes adultes âgés de 20 à 24 ans présentent le degré de mobilité géographique le plus élevé, un tiers d'entre eux ayant déménagé entre mars 1993 et mars 1994. Ces données révèlent également des différences liées à la race, le taux de mobilité étant plus élevé parmi les Noirs et les Hispaniques (19,6 % et 22,4 %, respectivement), comparativement aux Blancs (16,0 %, cf. Hansen 1994). Enfin, il y a également une corrélation étroite entre le mode

Les données utilisées pour la présente analyse proviennent de la Living Situation Survey (LSS), enquête conçue spécialement pour recueillir des renseignements sur le nombre de personnes faisant partie des ménages, sur les liens sociaux, sur la mobilité et sur le lieu de résidence habituel. La LSS était une enquête à participation volontaire qui a été menée par le Research Triangle Institute (RTI) et

3. MÉTHODE

présentes lors de la visite du recenseur.

souvent limitée aux personnes qui sont effectivement couvertes censitaire dans le cas de ce genre de ménage est un peu partout aux États-Unis (Velasco 1992; Mahler 1993; Romero 1992). Ces chercheurs ont constaté que la couverture censitaire dans le cas de ce genre de ménage est souvent limitée aux personnes qui sont effectivement présentes lors de la visite du recenseur.

Comme dans le cas des ménages faisant partie d'un réseau de proches, le type de ménage décrit par Montoya comprend des personnes qui arrivent dans un ménage et qui le quittent; cependant, dans ce cas, les gens ont «des liens peu étroits entre eux ou pas de liens de parenté, et leurs séjours est éphémère», et souvent il s'agit de jeunes travailleurs migrants qui travaillent et dorment selon des horaires différents, et qui n'ont pratiquement pas de liens sociaux entre eux. Plusieurs autres ethnographes ont identifié des ménages similaires dans d'autres collectivités hispaniques, mentionnés comme faisant partie d'un ménage donné. En général, les membres de la périphérie fluide devraient être mentionnés comme faisant partie d'un ménage donné. En effet, les déplacements de ces personnes peuvent ne pas comporter un changement d'adresse permanente, ce qui peut rendre difficile la détermination du nombre de personnes qui vivent ou habitent à une adresse donnée.

La mobilité temporaire n'est pas un aspect que l'on trouve uniquement au sein de collectivités particulières. Il existe de nombreux exemples de ce type de mobilité dans la population en général, notamment une mobilité associée à des voyages d'affaires ou de vacances de longue durée, à la fréquentation d'universités, à des situations de garde et à des personnes qui sont présentes dans un ou plusieurs ménages durant une période donnée. Cette mobilité dans la périphérie fluide (mobilité temporaire) diffère de la mobilité géographique parce qu'elle consiste en des déplacements caractérisés par des départs d'un seule résidence qui sont généralement suivis d'un retour à celle-ci. Il est difficile pour les répondants d'un recensement ou d'une enquête de savoir quels membres de la périphérie fluide devraient être mentionnés comme faisant partie d'un ménage donné. En effet, les déplacements de ces personnes peuvent ne pas comporter un changement d'adresse permanente, ce qui peut rendre difficile la détermination du nombre de personnes qui vivent ou habitent à une adresse donnée.

Enfin, Montoya (1992) décrit un type de ménage très différent qui est typique de certaines collectivités d'immigrants hispaniques qui se sont formées ces dernières années. Comme dans le cas des ménages faisant partie d'un réseau de proches, le type de ménage décrit par Montoya comprend des personnes qui arrivent dans un ménage et qui le quittent; cependant, dans ce cas, les gens ont «des liens peu étroits entre eux ou pas de liens de parenté, et leurs séjours est éphémère», et souvent il s'agit de jeunes travailleurs migrants qui travaillent et dorment selon des horaires différents, et qui n'ont pratiquement pas de liens sociaux entre eux. Plusieurs autres ethnographes ont identifié des ménages similaires dans d'autres collectivités hispaniques, mentionnés comme faisant partie d'un ménage donné. En général, les membres de la périphérie fluide devraient être mentionnés comme faisant partie d'un ménage donné. En effet, les déplacements de ces personnes peuvent ne pas comporter un changement d'adresse permanente, ce qui peut rendre difficile la détermination du nombre de personnes qui vivent ou habitent à une adresse donnée.

Le genre de mobilité qui nous intéresse pourrait être sous-dénoté.

Le genre de mobilité qui nous intéresse pourrait être sous-dénoté.

Mobilité temporaire et déclaration du lieu de résidence habituel

NANCY BATES et ELEANOR R. GERBER¹

RÉSUMÉ

On émet l'hypothèse que la mobilité temporaire contribue à l'erreur de couverture au sein des ménages, car cette mobilité pourrait avoir une incidence sur la détermination du «lieu de résidence habituel», notion qui est couramment appliquée lorsqu'on dresse une liste de personnes dans le cadre d'enquêtes ou de recensements portant sur les ménages. Dans la présente communication, nous analysons divers types de mobilité temporaire, ainsi que les rapports existant entre ceux-ci et la détermination du lieu de résidence habituel. La mobilité temporaire est définie par le type de déplacement effectué à partir d'un seul lieu de résidence (où normalement on revient), au cours d'une période de référence de deux à trois mois. La typologie est établie à partir de deux aspects: la diversité des endroits visités et la fréquence des visites. À l'aide de données tirées de la U.S. Living Situation Survey (LSS), qui a été effectuée en 1993, on a identifié quatre types de mobilité temporaire. Dans le cas de deux catégories, on a constaté un comportement associé à des visites répétées et la présence d'un plus grand nombre de personnes qui ont tendance à échapper au dénombrement lors d'enquêtes et de recensements. La modélisation log-linéaire indique que les types de mobilité temporaire constituent une variable prédictive importante liée à la détermination du lieu de résidence habituel, même lorsqu'on tient compte du temps passé loin du ménage et des caractéristiques démographiques.

MOTS CLÉS: Mobilité temporaire; lieu de résidence habituel; listes de membres de ménages; couverture.

1. INTRODUCTION

Lors de tout recensement démographique, le défi principal consiste à obtenir un dénombrement exact et complet de toutes les personnes faisant partie de la population visée par le recensement en question. Par conséquent, on pourrait dire que le nombre de personnes qui échappent au dénombrement est l'indicateur le plus important d'après lequel on évalue la qualité d'un recensement. La plupart des enquêtes et des recensements fondés sur les ménages commencent par une question accompagnée d'une liste servant à indiquer tous les «résidents habituels» d'un ménage. Les études d'évaluation de la qualité des données de recensement indiquent que l'erreur de couverture constitue un problème. En 1990, la U.S. Post Enumeration Survey (PES) ainsi que des analyses démographiques ont estimé que le sous-dénombrement net à l'échelle nationale était d'environ 2 % (Hogan 1993; Robinson, Ahmed, Das Gupta et Woodrow 1993). D'autre études indiquent que l'erreur de couverture dans les enquêtes courantes (comme la U.S. Current Population Survey) est même plus importante que dans le cas des recensements décennaux (Shapiro, Diffendal et Cantor 1993; Chakrabarty 1992; Pennie 1990; Hainer, Hines, Martin et Shapiro 1988). Des travaux de recherche effectués par Fein et West (1988) ainsi que par Shapiro et coll. (1993) laissent penser que la sous-couverture d'un ménage est un élément plus important de l'erreur de couverture globale que le sous-dénombrement dû à la non-couverture d'un ménage complet. D'autres chercheurs indiquent que l'omission de membres d'un

ménage compte pour environ un tiers du sous-dénombrement global d'un recensement (Ellis 1994; Fay 1989a). La recherche portant sur la couverture indique également que les personnes qui ne sont pas dénombrées ne sont pas réparties au hasard au sein de la population. Ainsi, les Noirs et les Hispaniques font davantage l'objet d'un sous-dénombrement que les Blancs non hispaniques (4,6 % et 4,0 % respectivement, comparativement à 0,7 %; Hogan 1993). Les personnes qui habitent dans des structures à unités de logement multiples (comme des appartements) et celles qui sont locataires sont également plus susceptibles de ne pas figurer dans un dénombrement (Giffin et Mortality 1992; Mortality et Childers 1993; Ellis 1993). La présente communication est centrée sur un aspect que l'on soupçonne depuis longtemps être un facteur qui contribue à l'erreur de couverture au sein des ménages. Cet aspect a trait à la mobilité temporaire, c'est-à-dire au fait de quitter durant un certain temps un lieu de résidence. Nous examinons en particulier le nombre de lieux dans lesquels peut se rendre une personne, le nombre de ces déplacements et la durée de séjour en ces lieux. Nous cherchons également à savoir si la mobilité est un facteur qui a une incidence sur la couverture et si elle pourrait être un bon indicateur de l'attachement à un ménage. Nous formulons comme hypothèse que le niveau de mobilité d'une personne tend à avoir une incidence sur le choix qu'elle effectue le répondant d'un ménage quand il ou elle doit déterminer si la personne en question est un résident habituel ou non, et s'il faut la mentionner ou non dans une déclaration de recensement.

¹ Nancy Bates, Office of the Director, U.S. Bureau of the Census, Room 2031, Federal Building 3, Washington, DC 20233, et Eleanor R. Gerber, Center for Survey Methods Research, U.S. Bureau of the Census, Room 3133, Federal Building 4, Washington, DC 20233 U.S.A.

BIBLIOGRAPHIE

- ALHO, J.M. (1990). Logistic regression in capture-recapture models. *Biometrics*, 46, 623-635.
- ALHO, J.M., MÜLLER, M.H., WURDEMAN, K., et KIM, J. (1993). Estimating heterogeneity in the probabilities of enumeration for dual-system estimation. *Journal of the American Statistical Association*, 88, 1130-1136.
- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- CORMACK, R.M. (1989). Log-linear models for capture-recapture. *Biometrics*, 45, 395-413.
- DARROCH, J.N. (1958). The multiple-recapture census I: estimation of a closed population. *Biometrika*, 45, 343-359.
- DARROCH, J.N., FIENBERG, S.E., GLONEK, G.F.V., et JUNKER, B.W. (1993). A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability. *Journal of the American Statistical Association*, 88, 1137-1148.
- FAULKNERBERRY, G.D., et GAROU, A. (1991). Estimating a population total using an area frame. *Journal of the American Statistical Association*, 86, 445-449.
- FECISO, R., TORTORA, R.D., et VOGEL, F.A. (1986). Sampling frames for agriculture in the United States. *Journal of Official Statistics*, 2, 279-292.
- FIENBERG, S.E. (1972). The multiple recapture census for closed populations and incomplete 2^x contingency tables. *Biometrika*, 59, 591-603.
- HAINES, D.E. (1997). Estimating Population Parameters Using Multiple Frame and Capture-Recapture Methodology. Thèse de doctorat, North Carolina State University.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley & Sons.
- HARTLEY, H.O. (1962). Multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- HUGGINS, R.M. (1989). On the statistical analysis of capture experiments. *Biometrika*, 76, 133-140.
- KOTT, P.S., et VOGEL, F.A. (1995). Multiple-frame business surveys. *Business Survey Methods* (Ed., B.G. Cox). New York: John Wiley & Sons, 185-203.
- LUND, R.E. (1968). Estimators in multiple frame surveys. *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.
- NEALON, J.P. (1984). Review of the Multiple and Area Frame Estimators. Staff Report 80, U.S. Department of Agriculture, Statistical Reporting Service, Washington, DC.
- OTIS, D.L., BURNHAM, K.P., WHITE, G.C., et ANDERSON, D.R. (1978). Statistical inference for capture data on closed animal populations. *Wildlife Monographs*, 62, 1-135.
- POLLOCK, K.H., HINES, J.E., et NICHOLS, J.D. (1984). The use of auxiliary variables in capture-recapture and removal experiments. *Biometrics*, 40, 329-340.
- POLLOCK, K.H., TURNER, S.C., et BROWN, C.A. (1994). Techniques de saisie-ressaisie pour l'estimation de la taille de la population et de totaux de population lorsqu'on ne dispose pas d'une base de sondage complète. *Techniques d'enquête*, 20, 121-128.
- SANATHANAN, L. (1972). Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43, 1, 142-152.
- SEBER, G.A.F. (1982). *The Estimation of Animal Abundance and Related Parameters*. (2-ième Edition). New York: Macmillan.
- SEKAR, C.C., et DEMING, W.E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, 44, 101-115.
- SIRKEN, M.G. (1970). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- SKINNER, C.J. (1991). On the efficiency of taking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- SKINNER, C.J., HOLMES, D.J., et HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statistique*, 62, 333-347.
- WOLTER, K.M. (1990). Capture-recapture estimation in the presence of a known sex ratio. *Biometrics*, 46, 157-162.

Résultats des simulations pour $N = 5000$

P_B	θ	% Biases relatif	% PEQMR	% Biases relatif	% PEQMR	% Biases relatif	% PEQMR	P_A				
									0,5	0,10	0,20	
0,7	0,5 (0,462)	N_1	61,47	61,82	61,39	61,76	61,69	62,04				
		N_2	-0,18	15,78	0,26	10,65	-0,15	6,72				
		N_3	54,84	55,17	49,06	49,38	39,38	39,65				
		N_4	19,73	38,12	4,77	19,52	-0,01	7,21				
	1 (0,490)	N_1	-0,28	6,14	-0,13	5,99	0,35	6,15				
		N_2	0,43	18,14	0,47	12,85	-0,20	8,34				
		N_3	-0,22	5,82	-0,03	5,35	0,16	4,88				
		N_4	0,26	9,82	-0,04	7,44	0,11	5,95				
	1,5 (0,508)	N_1	-36,21	36,68	-36,29	36,78	-35,90	36,38				
		N_2	0,41	20,39	-0,16	14,21	0,39	9,55				
		N_3	-32,87	33,37	-29,97	30,49	-24,13	24,66				
		N_4	-19,11	31,15	-11,51	23,92	-3,12	14,03				
	2 (0,522)	N_1	-61,04	61,30	-60,53	60,81	-60,64	60,92				
		N_2	0,40	20,09	0,60	15,43	0,31	9,67				
		N_3	-55,69	55,96	-50,24	50,55	-41,46	41,76				
		N_4	-14,10	36,31	-2,34	20,96	0,26	9,84				
0,9	0,5 (0,806)	N_1	5,56	5,70	5,52	5,67	5,54	5,68				
		N_2	-0,12	4,55	0,11	3,19	-0,03	2,08				
		N_3	5,21	5,35	4,86	5,01	4,22	4,35				
		N_4	4,97	5,41	3,64	4,88	2,26	3,79				
	1 (0,810)	N_1	-0,02	1,58	0,08	1,55	0,01	1,57				
		N_2	-0,09	6,16	-0,17	4,08	-0,14	2,79				
		N_3	-0,03	1,53	0,05	1,48	-0,02	1,35				
		N_4	0,37	3,19	0,11	2,18	0,09	1,89				
	1,5 (0,814)	N_1	-4,66	5,00	-4,52	4,85	-4,61	4,90				
		N_2	-0,25	7,54	0,11	4,95	-0,09	3,14				
		N_3	-4,39	4,73	-3,96	4,32	-3,55	3,85				
		N_4	-2,50	6,31	-2,26	5,02	-1,84	3,82				
	2 (0,817)	N_1	-8,45	8,68	-8,38	8,60	-8,46	8,69				
		N_2	-0,21	7,86	-0,06	5,29	0,01	3,73				
		N_3	-7,95	8,18	-7,39	7,61	-6,49	6,73				
		N_4	-3,76	8,80	-2,77	6,99	-1,25	4,97				

Tableau 3

Tableau 2
Résultats des simulations pour $N = 500$

P_B	θ	% Biais relatif	% PEQMR	% Biais relatif	% PEQMR	% Biais relatif	% PEQMR	% Biais relatif	% PEQMR	P_A
0,7	0,5 (0,462)	N_1	62,30	66,01	60,64	64,04	63,26	66,81	22,58	0,20
		N_2	0,30	49,07	-0,75	32,37	0,85	43,32	22,58	
		N_3	55,52	58,95	48,15	51,15	40,53	43,32	43,32	
		N_4	48,15	58,88	37,88	49,25	24,95	38,80	38,80	
	1 (0,490)	N_1	0,47	19,26	1,01	19,08	-0,11	19,45	19,45	
		N_2	0,45	57,34	0,34	39,61	0,88	27,25	27,25	
		N_3	0,43	18,21	0,83	16,93	0,14	15,75	15,75	
		N_4	2,40	27,57	1,39	22,94	0,29	17,96	17,96	
	1,5 (0,508)	N_1	-35,60	40,06	-36,48	40,58	-35,69	40,26	40,26	
		N_2	3,11	66,43	-5,08	41,96	0,30	28,79	28,79	
		N_3	-32,07	36,79	-31,01	35,28	-24,04	28,88	28,88	
		N_4	-22,74	47,62	-26,21	37,57	-17,06	30,38	30,38	
	2 (0,522)	N_1	-60,07	62,91	-61,31	64,06	-60,41	63,28	63,28	
		N_2	-6,12	66,59	-1,15	46,68	1,67	30,99	30,99	
		N_3	-55,36	58,35	-51,21	54,19	-40,89	43,99	43,99	
		N_4	-41,39	63,79	-34,79	55,45	-18,60	41,35	41,35	
0,9	0,5 (0,806)	N_1	5,37	6,79	5,27	6,63	5,59	6,97	6,97	0,10
		N_2	0,08	14,78	-0,06	10,17	-0,06	6,55	6,55	
		N_3	5,04	6,44	4,62	5,93	4,24	5,53	5,53	
		N_4	5,94	9,48	5,03	7,05	4,34	5,72	5,72	
	1 (0,810)	N_1	0,30	5,01	0,17	5,01	0,25	4,94	4,94	
		N_2	0,78	20,72	0,41	14,06	-0,06	9,03	9,03	
		N_3	0,33	4,83	0,20	4,68	0,17	4,24	4,24	
		N_4	3,23	13,79	1,88	9,35	1,00	5,98	5,98	
	1,5 (0,814)	N_1	-4,29	7,07	-4,39	7,32	-4,55	7,37	7,37	
		N_2	-0,65	21,52	0,35	15,88	0,02	10,27	10,27	
		N_3	-4,07	6,78	-3,83	6,73	-3,49	6,15	6,15	
		N_4	-0,43	13,77	-1,18	10,92	-1,43	8,20	8,20	
	2 (0,817)	N_1	-8,28	10,27	-8,40	10,36	-8,33	10,32	10,32	
		N_2	-0,29	25,59	0,39	17,66	0,35	11,41	11,41	
		N_3	-7,80	9,82	-7,35	9,38	-6,30	8,20	8,20	
		N_4	-2,52	17,96	-3,10	14,02	-2,73	10,33	10,33	

également deux méthodes de stratification qui sont utiles lorsque la stratification d'une base areolaire et des listes se fait en fonction de la même variable. Ces résultats feront l'objet de futures publications.

6. DISCUSSION

Cette étude porte principalement sur l'estimation de la taille de la population, à partir de plusieurs bases de sondage. L'information provenant d'une base areolaire et/ou d'une ou de plusieurs listes est recueillie et combinée, pour obtenir divers estimateurs. Nous calculons les estimateurs de la taille de la population, lorsque l'information est disponible uniquement pour k listes indépendantes et aussi lorsque l'information est disponible pour un échantillon de la base areolaire, en plus des listes. Nous présentons ensuite une étude de simulation, qui vise à comparer la performance des estimateurs dans le cas spécial où l'on utilise deux listes et une base areolaire. À la lumière des résultats de cette simulation, nous recommandons l'estimateur calculé à partir de la fonction de vraisemblance indépendante intégrale, N_3 , lorsque les listes sont indépendantes ou presque indépendantes. Dans le cas de dépendance de modérée à forte, nous recommandons plutôt l'estimateur de sélection N_2 .

Nous examinons aussi l'estimation des chiffres de la population, en regard de deux scénarios. Dans le premier cas, nous supposons que des observations sont disponibles pour toutes les unités qui forment les listes. Dans le deuxième cas, par contre, nous supposons que l'information n'est disponible que pour des sous-échantillons de chaque liste. Nous utilisons un estimateur de type Horvitz-Thompson lorsque les listes sont indépendantes et un estimateur de sélection, lorsque les listes sont dépendantes. Dans cette étude, nous nous sommes intéressés principalement à l'estimation de la taille de la population. En pratique, toutefois, il se pourrait que l'on veuille estimer les chiffres de la population en regard de plusieurs caractéristiques, selon un plan d'échantillonnage à plusieurs degrés avec probabilités d'inclusion inégales. Au nombre des études qui traitent de ce dernier sujet, mentionnons celles de Bankier (1986), Skinner (1991) et Skinner, Holmes et Holt (1994).

7. REMERCIEMENTS

Les auteurs aimeraient remercier le rédacteur et les deux examinateurs pour leurs commentaires utiles au sujet d'une version antérieure du présent article. Cette recherche a été financée partiellement par le U.S. Geological Survey, Biological Resources Division. Christine Bunck est directrice du programme BEST. Les vues qui y sont exprimées appartiennent aux auteurs et ne sont pas nécessairement partagées par le Census Bureau.

L'information extraite de la base areolaire augmente, ces pourcentages diminuent. De même, à mesure que la taille de la population augmente, de 500 à 5 000, les pourcentages d'erreur quadratique moyenne relative diminuent eux aussi. Comme les valeurs de p_d dans notre étude de simulation sont faibles, la variance de N_2 est élevée. Cependant, même si N_3 comporte un biais, son erreur-type est très faible et ceci se traduit par un pourcentage d'erreur quadratique moyenne relative plus faible. L'estimateur N_4 réduit le biais de N_3 mais comporte une grande erreur-type; N_4 n'est donc pas un estimateur particulièrement utile. Lorsque les valeurs de θ et p_d sont plus élevées, N_2 devrait donner de meilleurs résultats que N_3 . Pour leurs valeurs de θ et p_d examinées ici, nous recommandons d'utiliser l'estimateur N_3 , de préférence à tous les autres proposés.

5.4 Limitations de l'étude

Notre étude avait pour but de comparer le biais, l'erreur-type et l'erreur quadratique moyenne de quatre estimateurs de la taille de la population, en présupant de probabilités d'inclusion égales dans les deux listes. De futures études pourraient être menées en incluant des probabilités d'inclusion inégales et des valeurs de θ plus élevées. De toute évidence, l'avantage de N_3 sur N_1 dépend du coût d'échantillonnage d'une base areolaire. Pour notre étude, nous avons retenu uniquement les cas où les valeurs de p_d étaient faibles. Or de faibles valeurs de p_d sont associées à des coûts élevés d'échantillonnage par base areolaire. Cependant, même dans ce cas, nous observons une réduction significative des pourcentages de l'erreur quadratique moyenne relative et du biais relatif, ce qui justifie l'utilisation de N_3 sur N_1 . Nous n'examinons pas ici de fonction objective qui tiendrait compte à la fois des coûts d'échantillonnage et des pourcentages de l'erreur quadratique moyenne relative et du biais relatif.

Tout au long de cet article, nous avons présupé que la probabilité d'inclusion, à l'intérieur d'une liste donnée, était la même pour toutes les unités. Haines (1997) examine le cas où les probabilités d'inclusion sont représentées comme étant une fonction d'une covariable. Lorsque les probabilités d'inclusion sont hétérogènes, il se peut que la probabilité d'inclusion dans une liste soit alors plus élevée pour les grandes unités que pour les petites. Les probabilités d'inclusion hétérogènes jouent un rôle important dans l'estimation des chiffres de la population, lorsque la variable de réponse présente une distribution fortement asymétrique ou des valeurs rares. Haines (1997) propose

Pour chaque combinaison paramétrique, nous produisons les données $(n_a, n_{b_1}, n_{b_2}, n_{ab_1}, n_{ab_2}, n_{b_1b_2}, n_{ab_1b_2})$. Un millier de répétitions de Monte Carlo sont effectuées pour chaque combinaison paramétrique.

5.2 Estimateurs

Nous comparons quatre estimateurs de la taille de la population, $\hat{N}_1, \hat{N}_2, \hat{N}_3$, et \hat{N}_4 . \hat{N}_1 est l'estimateur de Lincoln-Petersen qui n'inclut pas d'information de la base areolaire. L'estimateur \hat{N}_1 convient lorsque les listes sont indépendantes. Comme cet estimateur ne tient pas compte de l'information de l'échantillon de la base areolaire, on s'attend à ce qu'il soit inefficace lorsque l'information de la base areolaire est disponible. L'estimateur de sélection, \hat{N}_2 , fait la somme des estimations par domaines de chevauchement et sans chevauchement et convient tout particulièrement dans les cas de listes dépendantes. Le troisième estimateur, \hat{N}_3 , est calculé à partir de la fonction de vraisemblance de la base de sondage indépendante intégrale. Cet estimateur s'appuie sur l'information extraite de la base areolaire et sur le fait que les listes sont

indépendantes ($\theta = 1$). Nous croyons que \hat{N}_3 est le meilleur estimateur lorsque les listes B_1 et B_2 sont indépendantes, tandis que \hat{N}_2 serait le meilleur lorsque il y a dépendance. Nous avons donc examiné également un estimateur d'avant essai pour tester l'indépendance des listes. Selon notre définition, \hat{N}_4 est égal à \hat{N}_2 lorsque les données portent fortement à croire que les listes B_1 et B_2 ne sont pas indépendantes. Sinon, $\hat{N}_4 = \hat{N}_3$. Officiellement,

$$\hat{N}_4 = \begin{cases} \hat{N}_2 & \text{si VDA} > \chi^2_{1,0.05} = 3,84 \\ \hat{N}_3 & \text{autrement,} \end{cases}$$

où VDA est la variable chi-carré du test de validité de l'ajustement pour tester $H_0: \theta = 1$ et est calculé à partir du tableau à double entrée suivant.

	Dans B_1	Pas dans B_1
Dans B_2	$n_{ab_1b_2}$	n_{ab_1}
Pas dans B_2	n_a	n_{aB_1}

Figure 1. Classification des éléments échantillonnés à partir de la base areolaire

La figure 1 répartit les n_a éléments selon qu'ils sont présents dans les listes B_1 et B_2 ou qu'ils en sont absents.

5.3 Comparaison des estimateurs

Les tableaux 2 et 3 présentent les pourcentages du biais relatif et de l'erreur quadratique moyenne relative des estimateurs $\hat{N}_1, \hat{N}_2, \hat{N}_3$, et \hat{N}_4 pour des populations de tailles correspondant respectivement à 500 et 5 000. Nous

réduisons le biais et l'erreur quadratique moyenne par N afin de pouvoir comparer directement des estimateurs basés sur des populations de tailles différentes. Une comparaison entre \hat{N}_1 et \hat{N}_3 montre l'avantage qu'il y a à prélever un échantillon de la base areolaire. En pratique, ces avantages dépendent du coût relatif de l'échantillon de la base areolaire. Cependant, nous ne tenons pas compte, dans la présente étude, des coûts d'échantillonnage. La probabilité d'être inclus dans les deux listes, p_{11} , est indiquée entre parenthèses, sous la colonne θ . Lorsque $p_B = p_C = 0,9$, la valeur de p_{11} doit se situer entre 0,8 et 0,9. Cependant, lorsque θ varie de 0,5 à 2, p_{11} se situe uniquement entre 0,806 et 0,817.

L'estimateur \hat{N}_2 est sans biais pour N et a le plus faible pourcentage de biais relatif. Les estimateurs \hat{N}_1 et \hat{N}_3 sont asymptomatiquement cohérents pour N et donnent des biais dont la valeur se rapproche de 0, lorsque $\theta = 1$. Par contre, \hat{N}_1 et \hat{N}_3 ont un large biais lorsque $\theta \neq 1$. Le biais relatif, en pourcentage, de \hat{N}_4 est inférieur à celui de \hat{N}_3 mais il ne se rapproche pas de zéro. Le biais ne change pas de façon significative à mesure que p_a augmente de 0,05 à 0,10 à 0,20.

Lorsque $N = 500$ et $p_B = p_C = 0,9$, \hat{N}_3 a le plus faible pourcentage d'erreur quadratique moyenne relative. Ceci s'explique notamment du fait que l'éventail limité des valeurs de p_{11} est similaire à la valeur de p_{11} lorsque il y a indépendance (0,810). Le pourcentage de l'erreur quadratique moyenne relative pour \hat{N}_3 est de 40 à 50 % inférieur à celui de \hat{N}_2 . Par contre, il n'est que de 15 à 30 % inférieur à celui de \hat{N}_1 . Par conséquent, lorsque la probabilité d'inclusion dans les listes est très élevée, alors \hat{N}_1 et \hat{N}_3 sont tous deux nettement préférables à \hat{N}_2 . En outre, si les coûts d'échantillonnage dans la base areolaire sont élevés, alors \hat{N}_1 peut s'avérer un estimateur de remplacement acceptable pour \hat{N}_3 . Lorsque $N = 500$ et $p_B = p_C = 0,7$, c'est \hat{N}_3 qui a la plus faible erreur quadratique moyenne relative, en pourcentage, lorsqu'il y a indépendance; lorsque $\theta = 2$, c'est \hat{N}_2 qui a le plus faible pourcentage d'erreur quadratique moyenne relative. Si $N = 5 000$ et $p_B = 0,7$, alors \hat{N}_3 a le plus faible pourcentage d'erreur quadratique moyenne relative, seulement lorsque $\theta = 1$. Pour toutes les autres valeurs de θ , c'est \hat{N}_2 qui obtient le plus faible pourcentage. Dans tous les cas, \hat{N}_3 présente une très faible variance et l'erreur quadratique moyenne relative, exprimée en pourcentage, est due principalement au biais dans \hat{N}_3 . Pour $\theta < 1$, \hat{N}_3 tend à avoir un biais positif, alors que \hat{N}_3 a un biais négatif lorsque $\theta > 1$. Dans le cas où $N = 5 000$ et $p_B = 0,9$, \hat{N}_3 le plus faible pourcentage d'erreur quadratique moyenne relative lorsque $\theta = 1$, mais c'est \hat{N}_2 qui a le plus faible pourcentage lorsque $\theta = 0,5$ et 2. Enfin, lorsque $\theta = 1,5$, aucun estimateur n'est supérieur à un autre, en ce qui a trait au pourcentage d'erreur quadratique moyenne relative.

Comme prévu, les pourcentages d'erreur quadratique moyenne relative de \hat{N}_2, \hat{N}_3 , et \hat{N}_4 diminuent à mesure que la valeur de p_a augmente. Par conséquent, à mesure que

4. LISTES DÉPENDANTES

Examinons maintenant le cas où il y a dépendance entre les listes, mais où la base areolaire et les listes demeurent indépendantes. Par exemple, dans les expériences par capture et recapture, la probabilité qu'un animal soit capturé au deuxième échantillonnage peut dépendre de sa capture au premier échantillonnage. Voir Fienberg (1972), Cormack (1989), Wolter (1990), Pollock, Hines et Nichols (1984), Huggins (1989) et Alho (1990) pour obtenir des exemples précis.

Prenons le cas où nous avons deux listes, B_1 et B_2 , qui sont dépendantes. Supposons que P_{11} représente la probabilité d'être inclus dans les deux listes. Si B_1 et B_2 sont indépendantes, alors $P_{11} = P_{B_1} P_{B_2}$ où P_{B_1} et P_{B_2} sont les probabilités d'inclusion, respectivement pour B_1 et B_2 . Supposons également que P_{10} (P_{01}) est la probabilité d'être inclus dans la liste B_1 (B_2) mais non dans la liste B_2 (B_1). La probabilité d'exclusion des deux listes est représentée par $P_{00} = 1 - P_{B_1} - P_{B_2} + P_{11}$.

La fonction de vraisemblance est définie par l'équation

$$\mathcal{L}(P_{B_1}, P_{B_2}, P_{11}, N | P^A, n^a, N_{B_1}, n_{B_1}, N_{B_2}, n_{B_2}, N_{ab_1}, n_{ab_1}, N_{ab_2}, n_{ab_2}, N_{ab_1b_2}, n_{ab_1b_2}) = \binom{N}{n^a, N_{B_1}, N_{B_2}, n_{ab_1}, n_{ab_2}, N_{B_1B_2}, n_{ab_1b_2}} P_{11}^{n_{ab_1b_2}} P_{B_2}^{n_{B_2}} (1 - P^A)^{N - n^a}$$

$(P_{B_1} - P_{11})^{N_{B_1} + n_{ab_1}} (P_{B_2} - P_{11})^{N_{B_2} + n_{ab_2}} P_{11}^{N_{B_1B_2} + n_{ab_1b_2}}$
En maximisant la valeur de (12) par rapport à P_{B_1}, P_{B_2}, P_{11} et N , on obtient l'approximation suivante

$$\hat{N} = N_{B_1} + N_{B_2} + n_{ab_1} + n_{ab_2} + N_{B_1B_2} + n_{ab_1b_2} + \frac{n^a}{n^a + n_{ab_1b_2}}$$

qui coïncide avec l'estimateur de sélection \hat{N}_2 . En d'autres mots, \hat{N} est également l'estimateur qui est obtenu en combinant les deux listes en une seule, où les doublements sont éliminés et où la taille du domaine sans chevauchement est estimée à partir de l'échantillon de la base areolaire. Il peut également être démontré que la méthode du maximum de vraisemblance à deux degrés de Sanathanan (1972) mène à

$$\hat{N} = \frac{n^a + N_{B_1 \cup B_2}}{N_{B_1 \cup B_2}} \frac{\hat{N}_2}{P^A + (1 - P^A) \frac{\hat{N}_2}{N_{B_1 \cup B_2}}}$$

Par conséquent, l'estimateur du maximum de vraisemblance et l'estimateur de Sanathanan coïncident tous deux avec l'estimateur de sélection. Si l'on possède des données provenant de deux listes dépendantes mais que la nature du lien de dépendance est inconnue, alors nous ne pouvons estimer les paramètres individuels. Lorsque l'information

5. ÉTUDE DE SIMULATION

d'une base areolaire indépendante est disponible, tous les paramètres sont estimables. Cependant, pour estimer N , il suffit d'avoir $N_{B_1 \cup B_2}$ aucune information additionnelle ne nous est donnée par N_{B_1}, N_{B_2} , et $N_{B_1B_2}$. Il existe différentes méthodes pour modéliser la dépendance entre k listes, pour estimer la taille et les chiffres de la population. Des informations additionnelles sur la population ou de l'information provenant d'une base areolaire indépendante sont nécessaires pour modéliser avec précision la dépendance. Fienberg (1972) et Cormack (1989) proposent des modèles loglinéaires contraints pour modéliser la dépendance. Pour sa part, Wolter (1990) utilise des contraintes externes, comme un rapport de masculinité connu, pour estimer la taille de la population lorsqu'il y a dépendance. Une autre technique consiste à modéliser les probabilités d'inclusion comme étant une fonction des covariables. Alho, Mulry, Wurdeman et Kim (1993) utilisent un modèle de régression logistique conditionnel pour estimer la probabilité d'être dénombré lors d'un recensement et appliquent ce modèle à l'enquête postcensitaire de 1990. Le rôle des variables auxiliaires dans les expériences par capture et recapture avec probabilités inégales est examiné par Pollock et coll. (1984), Huggins (1989) et Alho (1990).

5.1 Conception de l'étude

Afin d'étudier à la fois les cas dépendants et indépendants, nous définissons le paramètre θ qui reflète la structure de dépendance entre les listes B_1 et B_2 . Ce paramètre a la même forme que le rapport de cotes et il est représenté officiellement par l'équation

$$\theta = \frac{P_{00}P_{11}}{P_{01}P_{10}}$$

Dans le cas des deux listes, la valeur de θ détermine une solution unique pour P_{11} . Dans notre étude, les facteurs

varient comme suit:

Facteur	Niveau	Définition
N	500, 5 000	Taille de la population
P^A	0,05, 0,10, 0,20	Probabilité d'inclusion pour la base areolaire A
$P_{B_1} (= P_{B_2})$	0,7, 0,9	Probabilité d'inclusion pour la liste B_1 (B_2)
θ	0,5, 1,0, 1,5, 2,0	Rapport de cotes

d'éléments dans la base B_1 et la valeur N_{B_1} . Le deuxième groupe renferme les éléments de la base aréolaire qui ne sont pas inclus dans la ou les listes; il est donc désigné domaine sans chevauchement. La taille de ce dernier domaine correspond à une quantité aléatoire non observée, N_a . Le terme n_a désigne le nombre d'éléments que l'on trouve dans les n_j segments qui ne sont pas inclus dans la ou les listes, en vertu d'une règle d'association précise. Une valeur estimée de N_a est n_a/p_j . Par conséquent, \hat{N}_j à l'équation (9), fournit une estimation de la taille de la population. L'EMV de p_{B_1} qui en résulte est

$$\hat{p}_{B_1} = \frac{N_{B_1}}{N_a + \frac{N_a}{p_j}}.$$

Lorsque des listes multiples sont disponibles, il est possible de les combiner en une seule liste et d'utiliser l'estimateur qui précède pour obtenir une valeur estimée de N . En d'autres mots, supposons que nous avons l'estimateur de sélection

$$\hat{N}^2 = \hat{N} = N_{B_1 \cup \dots \cup B_k} = \frac{N_a}{p_a} = N_{B_1} + \dots + N_{B_k} +$$

$$N_{B_1 B_2} + \dots + N_{B_1 \dots B_k} + \frac{N_a}{p_a}. \quad (10)$$

À noter que l'estimateur de sélection \hat{N}_2 convient, même lorsque les listes ne sont pas indépendantes les unes des autres. Nous discutons de ce point plus en détails à la section 4. L'utilisation de cette méthode pour une base aréolaire et deux listes indépendantes donne la fonction de vraisemblance

$$\mathcal{L}(p_{B_1}, p_{B_2}, N | p_a, n_a, N_{B_1}, N_{B_2}, n_{B_1 B_2}, n_{a B_1}, n_{a B_2}, n_{a B_1 B_2}) = \frac{N}{N_{B_1} N_{B_2}} \left(n_a, N_{B_1}, N_{B_2}, n_{a B_1}, n_{a B_2}, n_{a B_1 B_2} \right) \left(1 - p_a \right)^{N - n_a} \left(1 - p_{B_1} \right)^{N_{B_1} - n_{B_1}} \left(1 - p_{B_2} \right)^{N_{B_2} - n_{B_2}}.$$

L'EMV de N est

$$\hat{N}_2 = \hat{N} = (2p_a)^{-1} *.$$

$$\sqrt{\frac{(N_{B_1} + N_{B_2}) p_a (N_{B_1} - n_{B_1} - N_{B_2} + n_{B_1 B_2})^2}{(2p_a)^{-1} (N_{B_1} + N_{B_2}) p_a (N_{B_1} - n_{B_1} - N_{B_2} + n_{B_1 B_2})^2}} \quad (11)$$

où $n_{a B_1 B_2}$ représente le nombre d'éléments inclus dans les n_j segments de la région échantillonnée qui appartiennent aux deux listes. On peut obtenir une estimation de la variance de \hat{N}_2 en ayant recours à l'approximation par série

3.2 Estimation des chiffres de la population

Lorsque les valeurs de y_i sont connues pour tous les éléments dans les k listes indépendantes ainsi que pour un échantillon de segments d'une base aréolaire, nous utilisons un estimateur de Horvitz-Thompson pour estimer les chiffres de population. Rappelons-nous les hypothèses formulées:

1. La probabilité qu'une unité soit incluse dans la i -ième liste, p_{B_i} , est égale pour toutes les unités.
2. L'inclusion d'une unité dans une base est indépendante de son inclusion dans une autre base.
3. La probabilité qu'une unité soit incluse dans l'échantillon de la base aréolaire formé de n_j segments correspond à $p_j = n_j/U_j$.

Comme nous supposons que les unités de la population n'appartiennent qu'à un segment de la région et que toutes les unités à l'intérieur d'un segment échantillonné sont observées, la troisième hypothèse est valide. Par conséquent, la probabilité que le i -ième élément soit inclus dans au moins une des k listes ou dans l'échantillon de la base aréolaire, ou les deux, est

$$\pi_1 = 1 - (1 - p_a)(1 - p_{B_1})(1 - p_{B_2}) \dots (1 - p_{B_k}) =$$

$$\frac{\hat{N}}{n_a + n_{a B_1} + \dots + n_{a B_1 \dots B_k}}.$$

L'estimateur de Horvitz-Thompson pour les chiffres de la population est

$$\hat{Y}_L^{H-T} = \frac{n_a + n_{a B_1} + \dots + n_{a B_1 \dots B_k}}{\hat{N}} \sum_{i \in \text{échantillon}} y_i = \hat{Y}_L,$$

où \hat{Y}_L est la moyenne des éléments distincts dans les listes B_1, \dots, B_k et des éléments dans l'échantillon de la base aréolaire.

Nous pouvons également utiliser l'estimateur de sélection pour estimer les chiffres de population. Le total du domaine de chevauchement connu est combiné à un estimateur du total du domaine sans chevauchement (NOL) pour donner $\hat{Y}_S = Y_L + \sum_{i \in \text{NOL}} y_i / p_a$. Le domaine sans chevauchement est formé des éléments de la base aréolaire qui ne figurent dans aucune liste et $Y_L = X_{B_1 \cup \dots \cup B_k}$ est le total des unités distinctes dans les k listes. Dans le cas du sous-échantillonnage, nous pouvons remplacer Y_L dans \hat{Y}_S par l'estimateur de Lund, représenté par

$$\hat{Y}_L^{L} = N_{B_1} \bar{y}_{B_1} + \dots + N_{B_k} \bar{y}_{B_k} + N_{B_1 B_2} \bar{y}_{B_1 B_2} + \dots + N_{B_1 \dots B_k} \bar{y}_{B_1 \dots B_k}.$$

estimé de Horvitz-Thompson coïncide avec l'estimateur des chiffres de la population proposé par Pollock, Turner et

Brown (1994).

Dans certains cas, les valeurs de la variable d'intérêt, y_i ,

ne sont pas disponibles pour toutes les unités dans les listes.

Si les listes sont grandes, des échantillons aléatoires sont alors prélevés de chaque liste et des données sont recueillies

sur ces sous-échantillons. S'il y a k listes, il est possible de

définir 2^k domaines. Nous examinons un prolongement de

l'estimateur proposé par Lund (1968) pour le total de toutes

les unités dans les listes

$$\hat{Y}_{L,L} = \sum_{i=1}^{2^k-1} N_i \bar{y}_i,$$

lequel est la somme pondérée de $2^k - 1$ moyennes du domaine, \bar{y}_i . Les facteurs de pondération sont déterminés

en fonction de la taille du domaine. L'estimateur des

chiffres de la population est

$$\hat{Y} = N \frac{\hat{Y}_{L,L}}{\sum_{i=1}^{2^k-1} N_i}.$$

3. LISTES MULTIPLES COMBINÉES À UNE BASE ARÉOLAIRE

3.1 Estimation de la taille de la population

Combiner de multiples listes individuelles à une base

aréolaire est une des solutions qui s'offrent pour pallier les

lacunes des listes d'échantillonnage. Supposons que la

région géographique qui nous intéresse est subdivisée en

U_j segments. Supposons également qu'un échantillon

aléatoire simple formé de n_j segments est sélectionné à

partir des U_j segments qui couvrent l'ensemble de la

population. Par conséquent, la probabilité qu'un segment

soit sélectionné correspond à $p_j = n_j/U_j$. Dans certaines

enquêtes, il est possible de subdiviser la région en segments

de taille à peu près égale. En pareils cas, la probabilité de

sélection d'un segment correspond à peu près à la

proportion de la région échantillonnée. L'inclusion d'une

base aréolaire ajoute à l'intégralité de la population-cible

(Hartley 1962). Nous supposons que chaque unité de

déclaration appartient à exactement un segment. Lorsqu'un

segment est sélectionné, toutes les unités de déclaration à

l'intérieur du segment sont observées. Pour estimer, par

exemple, le nombre de nids d'aigles à tête blanche, on

suppose que chaque nid n'appartient qu'à un segment et un

seul. Cependant, cette hypothèse n'est pas toujours valide.

Examinons par exemple le cas d'une exploitation porcine

qui s'étendrait au-delà des frontières du segment; dans un

tel cas, les éléments de la population peuvent être associés

à plus d'un segment. Pour résoudre ce problème, des règles

d'association établissant des liens entre les éléments de la

l'estimation. Voir Fraulkenberry et Garoui (1991) pour plus

de détails à ce sujet. Le National Agricultural Statistics

Service utilise trois règles de correspondance pour répartir

les éléments de la population entre les segments échantil-

lonnés. Les estimateurs de segment ouvert, fermé et pondéré

sont décrits dans Nealon (1984) et aussi dans Sirkén (1970).

Examinons le cas où nous avons k listes indépendantes

et une base aréolaire. La taille de la population, N , et les

probabilités d'inclusion dans la liste, p_{b_i} , $i = 1, \dots, k$, sont

des paramètres inconnus. Cependant, la probabilité

d'inclusion dans la base aréolaire $p_A = n_A/U_A$ est connue.

La fonction de vraisemblance est représentée par

$$\mathcal{L}(p_{b_1}, \dots, p_{b_k}, N | p_A, n_A, n_{ab_1}, \dots, n_{ab_1 \dots b_k}, N_{b_1}, \dots, N_{b_1 \dots b_k}) = \binom{N}{n_A, n_{ab_1}, \dots, n_{ab_1 \dots b_k}} \left(p_A^{n_A} (1 - p_A)^{N - n_A} \prod_{i=1}^k p_{b_i}^{N_{b_i}} (1 - p_{b_i})^{N - N_{b_i}} \right),$$

dans aucune autre liste.

Les EMV des paramètres sont représentés par $\hat{p}_{b_i} =$

N_{b_i}/N , où N est une solution au polynôme de k -ième degré

$$N(1 - p_A)(1 - \hat{p}_{b_1}) \dots (1 - \hat{p}_{b_k}) =$$

$$(N - n_A - n_{ab_1} - \dots - n_{ab_1 \dots b_k} - N_{b_1} - \dots - N_{b_1 \dots b_k}). \quad (8)$$

Des méthodes numériques sont essentielles pour résoudre l'équation (8) servant à calculer l'EMV \hat{N} de N . Parmi les

k racines de (8), nous sélectionnons \hat{N} qui maximise la

vraisemblance.

En appliquant cette méthode à une liste et une base

aréolaire, nous obtenons

$$\hat{N} = N_{b_1} + \frac{p_A}{n_A}. \quad (9)$$

Cet estimateur est également connu sous le nom d'estimateur de sélection (Kott et Vogel 1995), lequel répartit les éléments en deux groupes distincts. Le premier groupe renferme les éléments qui appartiennent à la fois à la liste et à la base aréolaire et il est désigné domaine de chevauchement. Comme on présume que tous les éléments dans une liste appartiennent à la base aréolaire, la taille du domaine de chevauchement coïncide avec le nombre

$$\frac{\mathcal{E}(N-1)}{\mathcal{E}(N)} = \frac{(\hat{N} - N^{b_1} - N^{b_2} - N^{b_1 b_2})}{\hat{N}} *$$

$$(3) \quad (1 - p_{B_1})(1 - p_{B_2}) = 1.$$

Nous présumons ici que N est large, de sorte que

$$\frac{N^{b_1}}{N} \approx \frac{N-1}{N^{B_1}} \quad \text{et} \quad \frac{N^{b_2}}{N} \approx \frac{N-1}{N^{B_2}}.$$

Si l'on remplace les estimateurs de (2) dans (3), on

obtient alors

$$(4) \quad \hat{N}_1 = \hat{N} = \frac{N^{B_1} N^{B_2}}{N^{b_1 b_2}}.$$

Sekar et Deming (1949) ont calculé une estimation de la variance de (4), exprimée par

$$V(\hat{N}_1) = \frac{(N^{b_1 b_2})^3}{N^{B_1} N^{B_2} N^{b_1} N^{b_2}}.$$

Le remplacement de (4) dans (2) donne les EMV de p_{B_1} et p_{B_2}

$$\hat{p}_{B_1} = \frac{N^{b_1}}{N^{b_1 b_2}} \quad \text{et} \quad \hat{p}_{B_2} = \frac{N^{b_2}}{N^{b_1 b_2}}.$$

L'estimateur \hat{N}_1 de N dans (4) est désigné estimateur de

Lincoln-Petersen, dans les modèles par capture et recapture à l'intérieur d'une population fermée. Les éléments de la liste B_1 peuvent être considérés comme les unités saisies lors du premier échantillonnage, alors que les éléments de la liste B_2 seraient les unités saisies au deuxième échantillonnage. Les éléments dans le domaine $b_1 b_2$ correspondent aux éléments saisis à la recapture. Étant donné cette correspondance, on constate facilement que la fonction de vraisemblance pour la taille de la population et les probabilités de capture, pour les deux échantillonnages, sera la même qu'en (1). Par conséquent, les EMV calculés pour deux listes indépendantes seront les mêmes que les EMV correspondants avec le modèle par capture et recapture avec deux échantillonnages.

Si nous poussons plus loin ces hypothèses, nous pouvons prétendre que le fait de combiner k listes indépendantes correspond directement au fait d'avoir k échantillonnages avec le modèle M_l selon des modèles par capture et recapture à l'intérieur d'une population fermée, où $t = k$ (Otis et coll. 1978). La fonction de vraisemblance générale pour k listes indépendantes, B_1, B_2, \dots, B_k , prend la forme

$$(5) \quad \mathcal{E}(p_{B_1}, \dots, p_{B_k} | N^{b_1}, \dots, N^{b_1 \dots b_k}) = \frac{N^{b_1 \dots b_k}}{N} \prod_{l=1}^k p_{B_l}^{N^{b_l}} (1 - p_{B_l})^{N - N^{b_l}}$$

2.2 Estimation des chiffres de population

Supposons que les valeurs mesurées y_l sont connues pour toutes les unités dans les k listes indépendantes. La probabilité estimée que le premier élément soit inclus dans au moins une des k listes est égale à

$$\pi_1 = P[B_l \cup_{l=1}^k B_l] = 1 - (1 - p_{B_1})(1 - p_{B_2}) \dots (1 - p_{B_k}),$$

où $\hat{p}_{B_l} = N_{B_l} / \hat{N}$ et \hat{N} est l'EMV de N calculé à partir de l'équation (7), laquelle équation (7)

$$\frac{\hat{N}}{(1 - \pi_1)} = 1 \quad \text{devenit, sous forme simplifiée,}$$

$$\hat{\pi}_1 = \frac{\hat{N}}{N^{b_1} + \dots + N^{b_1 \dots b_k}}.$$

Un estimateur de Horvitz-Thompson (1952) des chiffres de la population est exprimé par

$$\hat{Y}_{H-T} = \frac{1}{\sum_{y_l \in B_1 \cup \dots \cup B_k} \pi_1} = \frac{\hat{N}}{\sum_{y_l \in B_1 \cup \dots \cup B_k} y_l} = \hat{N} \hat{Y}_L,$$

où \hat{Y}_L est la moyenne des éléments distincts dans les listes. Par conséquent, pour k listes indépendantes, l'estimateur

uniquement à la base $B_1(B_2)$ et que le domaine b_1b_2 contient les unités $N_{b_1b_2}$ qui appartiennent aux deux bases de sondage. Le domaine final inclut les éléments de la population-cible existante qui ne figurent dans aucune des deux listes; sa taille correspond à $N - N_{b_1} - N_{b_2} - N_{b_1b_2}$. La notation du domaine pour les listes B_1 et B_2 est présentée au tableau 1. À noter que chaque élément dans chaque base doit être réparti dans un domaine, sans erreur. Les erreurs dans la détermination du domaine sont graves et ne peuvent être corrigées ultérieurement. Ces erreurs ne sont pas examinées durant la phase d'estimation et sont donc considérées comme des erreurs non dues à l'échantillonnage. Selon Nealon (1984), la détermination du domaine constitue la principale source d'erreur non due à l'échantillonnage dans un plan d'échantillonnage à bases multiples (Kott et Vogel 1995).

Tableau 1
Notation du domaine pour les listes B_1 et B_2

Taille du domaine	Probabilité du domaine
N_{b_1}	$p_{b_1} = p_{B_1}(1 - p_{B_2})$
N_{b_2}	$p_{b_2} = (1 - p_{B_1})p_{B_2}$
$N_{b_1b_2}$	$p_{b_1b_2} = p_{B_1}p_{B_2}$
$N - N_{b_1} - N_{b_2} - N_{b_1b_2}$	$1 - p_{b_1} - p_{b_2} - p_{b_1b_2} = (1 - p_{B_1})(1 - p_{B_2})$

Supposons que la probabilité qu'un élément de la population soit inclus dans la liste $B_1(B_2)$ est $p_{B_1}(p_{B_2})$. Comme on présume que les listes B_1 et B_2 sont indépendantes, la probabilité qu'un élément appartienne au domaine b_1 est $p_{b_1} = p_{B_1}(1 - p_{B_2})$. Les probabilités pour les autres domaines sont définies de la même manière. La taille de la population N et les probabilités d'inclusion p_{B_1} et p_{B_2} sont des paramètres inconnus. La fonction de vraisemblance est définie par l'équation

$$\mathcal{L}(p_{B_1}, p_{B_2}, N | N_{b_1}, N_{b_2}, N_{b_1b_2}) = \binom{N}{N_{b_1}, N_{b_2}, N_{b_1b_2}} p_{b_1}^{N_{b_1}} p_{b_2}^{N_{b_2}} p_{b_1b_2}^{N_{b_1b_2}} (1 - p_{b_1} - p_{b_2} - p_{b_1b_2})^{N - N_{b_1} - N_{b_2} - N_{b_1b_2}} \quad (1)$$

Les estimateurs du maximum de vraisemblance (EMV) des probabilités d'inclusion dans la base sont définis en maximisant le logarithme de la fonction de vraisemblance (1). Cette opération donne

$$\hat{p}_{B_1} = \frac{N_{b_1}}{N}, \quad \hat{p}_{B_2} = \frac{N_{b_2}}{N}, \quad \text{et} \quad \hat{p}_{B_1B_2} = \frac{N_{b_1b_2}}{N} \quad (2)$$

où l'EMV \hat{N} remplace N . Plutôt que de dériver le logarithme de la fonction de vraisemblance pour établir la valeur approximative de N , nous utilisons la «méthode du ratio» pour maximiser la vraisemblance où $\mathcal{L}(N)$ est égal à $\mathcal{L}(N - 1)$ (Dartoch 1958). Ce processus tient compte du paramètre discret N et donne l'équation

(1984) décrit en détails les estimateurs par bases multiples et base aréolaire qui sont utilisés par le ministère américain de l'Agriculture. Enfin, Kott et Vogel (1995) présentent une vue d'ensemble des enquêtes par bases multiples. Nous examinons, à la section 2, les estimations obtenues à partir d'information extraite de deux ou plusieurs listes indépendantes et démontrons le lien qui existe entre ces méthodes et celles par capture et recapture. À la section 3, nous examinons des estimateurs plus efficaces de la taille et des chiffres de la population, lorsque l'information d'une base aréolaire indépendante est disponible. Nous étendons ensuite ces méthodes aux listes dépendantes, à la section 4. Les résultats d'une étude de simulation qui compare différents estimateurs sont résumés à la section 5. Enfin, la section 6 présente un résumé de nos résultats et discute de futures orientations de recherche.

2. LISTES D'ÉCHANTILLONNAGE MULTIPLES

2.1 Estimation de la taille de la population

Les listes utilisées pour estimer la taille de la population sont habituellement incomplètes et ne couvrent pas l'ensemble de la population. Une solution pour pallier ce problème d'incomplétude est de fusionner deux ou plusieurs listes incomplètes. Le fait de combiner ainsi plusieurs listes peut améliorer la couverture de la population-cible et, de ce fait, fournir de meilleurs estimateurs. Dans le cas de listes multiples, on présume habituellement que la probabilité d'inclusion dans une liste donnée est égale pour chaque élément de la population; les éléments de la liste constituent donc eux-mêmes nos «échantillons». À titre d'exemple, il existe une probabilité égale que les gens choisissent ou non d'inscrire leur numéro de téléphone dans l'annuaire. Dans le cas des nids d'aigles à tête blanche, la liste de cette année est constituée à partir des nids observés l'année précédente. Si nous présumons que la probabilité qu'un nid soit dénombré est égale pour tous les nids, alors l'hypothèse qui précède est valide. L'hypothèse est également valide dans les expériences par capture et recapture où la première liste est formée de tous les animaux capturés au premier échantillonnage, alors que la deuxième contient tous les animaux capturés au deuxième échantillonnage. Ce scénario correspond au modèle M' dans les ouvrages traitant de l'échantillonnage par capture et recapture (voir par exemple Otis, Burnham, White et Anderson (1978) pour plus de détails à ce sujet). Le modèle M' suppose que le risque de capture, à chaque prélèvement, est le même pour tous les animaux dans la population; cette probabilité peut cependant varier d'un prélèvement à un autre. Examinons d'abord deux listes indépendantes, B_1 et B_2 . Supposons que B_1 compte N_{B_1} effectifs et que B_2 en compte N_{B_2} . Supposons également que le domaine $b_1(b_2)$ se compose des éléments $N_{b_1}(N_{b_2})$ qui appartiennent

Combinaison de bases multiples pour estimer la taille et les chiffres de la population

DAWN E. HAINES et KENNETH H. POLLOCK¹

RÉSUMÉ

Le présent article traite de méthodes efficaces d'estimation de la taille et des chiffres de la population, à partir de données extraites de listes multiples et d'une base aréolaire indépendante. Ces travaux constituent un prolongement de la méthode proposée par Hartley (1962), qui porte sur deux bases de sondage générales. Un des principaux inconvénients des listes vient de ce que celles-ci sont habituellement incomplètes. Nous proposons dans cet article plusieurs méthodes pour pallier ces lacunes. Un plan d'échantillonnage mixte alliant l'utilisation d'une liste et d'une base aréolaire permet d'inclure des bases de sondage multiples et de couvrir entièrement la population-cible. Pour chaque combinaison de bases de sondage qui est proposée, nous indiquons les notations qui s'y rapportent, la fonction de vraisemblance et les estimateurs de paramètres. Nous présentons également les résultats d'une étude de simulation qui compare les diverses caractéristiques des estimateurs proposés.

MOTS CLÉS : Base de sondage incomplète; échantillonnage par capture et recapture; estimateur de sélection; méthode à base double; estimation par bases multiples.

1. INTRODUCTION

La théorie classique de l'échantillonnage présume que la base de sondage est complète. En pratique, toutefois, cette hypothèse s'avère souvent non confirmée. En effet, les imperfections dans la base de sondage, dues par exemple à des omissions, des dédoublements et des enregistrements erronés, sont presque inévitables dans tout large exercice de collecte de données (Hansen, Hurwitz et Madow 1953). L'information recueillie à partir de listes et de bases aréolaires est utilisée pour estimer la taille et les chiffres d'une population inconnue. À titre d'exemple, un écologiste ou un biologiste de la faune peuvent utiliser une liste et une base aréolaire pour estimer le nombre de nids d'aigles à tête blanche à l'intérieur d'une région donnée. De même, le U.S. Bureau of the Census utilise une technique d'estimation double pour mesurer le sous-dénombrement du recensement décennal. Pour leur part, Darroch, Fienberg, Glonek et Jonker (1993) décrivent une méthode de saisie multiple avec trois échantillons pour estimer la taille de la population, lorsque les probabilités d'inclusion sont hétérogènes. Dans un même ordre d'idées, des autorités agricoles pourraient être intéressées à estimer par exemple le nombre d'exploitations porcines et le nombre total de porcs en Caroline du nord. Il est courant que des sources multiples soient utilisées pour estimer la taille et les chiffres de la population.

Les listes présentent une liste des unités d'échantillonnage dans la population-cible. Ces listes sont établies au fil des ans, à partir de l'information obtenue des scientifiques, des autorités municipales, des comités, des États et des organismes fédéraux. Les éléments d'information que

L'on retrouve sur une liste d'échantillonnage incluent par exemple le nom, l'adresse, le numéro de téléphone, le numéro de sécurité sociale ou la description physique du lieu. Ces éléments et d'autres variables de stratification de toutes sortes sont utilisés pour identifier des personnes, des animaux, des entreprises ou d'autres établissements. Pour estimer le nombre de nids d'aigles à tête blanche dans une région, la liste d'échantillonnage de l'année courante est construite à partir de celle de l'an dernier. Avec l'ajout des nouveaux nids, la liste de l'an dernier devient rapidement désuète et incomplète; en raison de cette incomplétude, les estimations basées uniquement sur les listes sous-estiment habituellement la taille réelle de la population. L'addition d'information complémentaire tirée d'une base aréolaire peut s'avérer une méthode efficace pour estimer la taille et les chiffres de la population.

Une base aréolaire est un ensemble de régions géographiques définies par des frontières identifiables. L'ensemble de la région dans laquelle les données sont recueillies est divisée en unités d'échantillonnage exhaustives et s'excluant mutuellement, désignées segments. Les segments sont habituellement stratifiés en fonction d'une caractéristique d'intérêt. Lorsqu'un échantillon aléatoire stratifié de segments a été prélevé, les enquêteurs visitent les segments échantillonnés et notent les mesures pour toutes les unités de déclaration qui s'y trouvent.

Le National Agricultural Statistics Service (NASS) utilise actuellement une méthode à bases multiples pour l'échantillonnage et l'estimation d'un grand nombre de denrées agricoles. Fecso, Tortora et Vogel (1986) pré-sentent une révision des bases d'échantillonnage utilisées pour le secteur agricole aux États-Unis, alors que Nealon

BIBLIOGRAPHIE

CASADY, R.J., et VALLIANT, R. (1993). Propriétés conditionnelles des estimateurs de stratification a posteriori selon la théorie normale. *Techniques d'enquête*, 19, 193-203.

COCHRAN, W.G. (1977). *Sampling Techniques*. New York: John Wiley and Sons.

DEVILLE, J.C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

ELTINGE, J.L., et JANG, D.S. (1996). Mesures de la stabilité des estimateurs des composantes de la variance dans un plan d'échantillonnage stratifié à plusieurs degrés. *Techniques d'enquête*, 22, 159-168.

ESTEVAO, V., HIDIROGLOU, M.A., et SÄRNDAAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

LEHTONEN, R., et PAHKINEN, E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. New York: Wiley.

MONTANARI, G.E. (1987). Post-sampling efficient QR-prediction in large-scale surveys. *Revue Internationale de Statistique*, 55, 191-202.

RAO, J.N.K. (1994). Estimating totals and distribution functions using auxiliary information a the estimation stage. *Journal of Official Statistics*, 10, 153-165.

SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.

WRIGHT, R.L. (1983). Finite population sampling with multivariate auxiliary information. *Journal of the American Statistical Association*, 78, 879-884.

REMERCIEMENTS

empiriques afin d'évaluer la stabilité quand la variable auxiliaire présente plusieurs dimensions et pour établir à quel moment la taille de l'échantillon permet de surmonter le problème.

Pour exploiter l'information fournie par $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$ dans la même enquête, on devra se pencher davantage sur les propriétés de distribution de cette statistique et de l'estimateur de régression assujéti à l'échantillon, qui semble donner de bons résultats dans l'étude empirique. Plus précisément, la distribution de $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$ quand le terme prend la valeur zéro facilitera le choix du seuil auquel passer de \hat{Y}_{r1} à \hat{Y}_{r2} devient vraiment intéressant. En plus d'accroître la taille de l'échantillon, on peut résoudre le problème de l'instabilité de la statistique en cherchant des estimateurs plus stables et plus convergents pour la variance et la covariance qu'on retrouve dans $\lambda(Y_{r1}, Y_{r2})$. Par ailleurs, puisqu'on s'intéresse à plus d'une variable dans la majorité des situations pratiques, il conviendrait de sélectionner l'estimateur optimal en fonction d'une mesure moyenne de λ pour les principales variables de l'enquête, afin d'appliquer les mêmes poids aux variables, la moyenne étant plus stable que les valeurs individuelles.

La présente recherche a été financée en partie grâce à une bourse du M.U.R.S.T. d'Italie. L'auteur remercie l'éditeur associé et les examinateurs pour leurs commentaires judicieux, qui ont considérablement amélioré la version initiale de l'article.

Tableau 3
Biais relatif (BR) empirique en pour cent et erreur quadratique moyenne (EQM) des estimateurs et proportion de l'échantillon pour $\hat{\lambda}(\hat{Y}_{r1k}, \hat{Y}_{r2k}) > 8\%$ dans la deuxième étude empirique

Variable auxiliaire	Estimateur	Taille de l'échantillon 40			Taille de l'échantillon 80		
		BR(%)	EQM	($\lambda > 8\%$)	BR(%)	EQM	($\lambda > 8\%$)
aucune	\bar{y}	0,01	100,0	—	0,01	100,0	—
(x)	\hat{Y}_{r11}	-0,01	55,2	82,6 %	0,00	54,3	85,0 %
(x)	\hat{Y}_{r31}	-0,05	48,4	—	-0,02	43,8	—
(1, x)'	\hat{Y}_{r12}	-0,01	51,7	72,7 %	0,00	50,8	83,2 %
(1, x)'	\hat{Y}_{r22}	-0,05	47,4	—	-0,01	43,3	—
(1, x)'	\hat{Y}_{r32}	-0,05	48,3	—	-0,02	43,8	—
(1, x)'	\hat{Y}_{r12}	0,02	51,6	—	0,01	50,7	—
(1, x)'	\hat{Y}_{r22}	0,02	44,3	—	0,00	42,3	—
(1, x, x ²)'	\hat{Y}_{r13}	-0,01	35,1	28,9 %	0,02	33,5	10,5 %
(1, x, x ²)'	\hat{Y}_{r23}	-0,10	38,0	—	-0,03	34,7	—
(1, x, x ²)'	\hat{Y}_{r33}	-0,04	37,0	—	-0,01	33,8	—
(1, x, x ²)'	\hat{Y}_{r13}	0,01	34,9	—	0,03	33,5	—
(1, x, x ²)'	\hat{Y}_{r23}	0,01	34,7	—	0,03	33,2	—

Tableau 4
Quelques propriétés des distributions empiriques de $\hat{\lambda}(\hat{Y}_{r1k}, \hat{Y}_{r2k})$, $k = 1, 2, 3$ (deuxième étude empirique)

Statistique	Taille de l'échantillon 40			Taille de l'échantillon 80		
	Moyenne	Ecart- type	Médiane	Quantiles 10 %	Quantiles 90 %	Quantiles 90 %
$\hat{\lambda}(\hat{Y}_{r11}, \hat{Y}_{r21})$	0,24	0,15	0,23	0,04	0,45	0,35
$\hat{\lambda}(\hat{Y}_{r12}, \hat{Y}_{r22})$	0,19	0,14	0,17	0,02	0,38	0,30
$\hat{\lambda}(\hat{Y}_{r13}, \hat{Y}_{r23})$	0,06	0,08	0,03	0,00	0,18	0,08

\hat{Y}_{r12} et \hat{Y}_{r22}). Le tableau 4 donne la moyenne, l'écart-type et certains quantiles de la distribution empirique de $\hat{\lambda}(\hat{Y}_{r1k}, \hat{Y}_{r2k})$, $k = 1, 2, 3$.

7. DISCUSSION

L'estimateur optimal peut s'avérer une solution efficace à l'estimateur de régression généralisé quand celui-ci s'articule sur des modèles de superpopulation mal spécifiés, pourvu que l'échantillon soit assez important. Cette efficacité peut être jaugée grâce à la statistique $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$, qui établit le gain asymptotique relatif d'efficacité de \hat{Y}_{r2} par rapport à \hat{Y}_{r1} pour l'échantillon, compte tenu d'une certaine somme d'informations auxiliaires. L'estimateur optimal semble donner de bons résultats, mêmes avec des échantillons de taille finie, et son emploi paraît utile, pourvu que la valeur de $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$ compense sa plus grande instabilité. En réalité, les résultats empiriques indiquent que l'estimateur optimal est plus instable, surtout avec une population asymétrique. On a besoin d'autres données

à \hat{Y}_{r11} et à \hat{Y}_{r12} d'extraire l'information de l'échantillon. D'un autre côté, \hat{Y}_{r13} est plus efficace que \hat{Y}_{r23} parce qu'il repose sur le modèle réel. La plupart du temps, $\hat{\lambda}(\hat{Y}_{r13}, \hat{Y}_{r23})$ se retrouve sous la valeur seuil, en particulier quand l'échantillon compte 80 éléments. L'estimateur \hat{Y}_{r33} assujéti à l'échantillon, est presque aussi efficace que \hat{Y}_{r13} . Si on examine les approximations linéaires, on remarque d'abord que l'EQM des estimateurs GREG \hat{Y}_{r12} et \hat{Y}_{r13} est égale presque toujours celle de \hat{Y}_{r12} et de \hat{Y}_{r13} , dans la deuxième étude. On ne peut en dire autant des estimateurs optimaux \hat{Y}_{r22} et \hat{Y}_{r23} . La diminution d'efficacité au niveau des approximations linéaires est plus importante pour \hat{Y}_{r22} et \hat{Y}_{r23} mais elle s'amenuise rapidement quand la taille de l'échantillon augmente. L'EQM des approximations linéaires confirme qu'avec une certaine somme de données auxiliaires, l'estimateur optimal entraîne une amélioration négligeable de l'efficacité, même avec les très gros échantillons (comparez \hat{Y}_{r13} et \hat{Y}_{r23}), quand le modèle sur lequel repose l'estimateur GREG se vérifie. On peut parvenir à des gains substantiels, si le modèle laisse à désirer, comme c'est le cas pour \hat{Y}_{r11} et \hat{Y}_{r12} (comparez

Tableau 2
Quelques propriétés des distributions empiriques de $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$ pour les populations gamma (première étude empirique)

Populations gamma	Moy- enne	Ecart- type	Médiane	Quantiles 10 % 90 %
$V(Y x) = 8x, n = 20$	10,7	9,8	8,7	1,3 24,9
$V(Y x) = 8x, n = 40$	9,2	6,3	8,3	2,5 19,1
$V(Y x) = 3x, n = 20$	21,6	12,3	19,2	6,9 40,7
$V(Y x) = 3x, n = 40$	19,0	9,5	18,9	9,4 34,2

La performance de \hat{Y}_{r3} ne manque pas non plus d'intérêt. Cet estimateur n'est presque pas biaisé et son EQM est plus faible que celle de \hat{Y}_{r1} , plus souvent on choisit \hat{Y}_{r2} . Le tableau 1 donne le pourcentage d'échantillons pour lesquels $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2}) > 8\%$ et \hat{Y}_{r2} a été sélectionné au lieu de \hat{Y}_{r1} pour chaque simulation. Plus la valeur théorique de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$, est élevée et plus souvent on retient \hat{Y}_{r2} au lieu de \hat{Y}_{r1} .

La performance de \hat{Y}_{r3} dépend manifestement de la distribution d'échantillonnage des statistiques $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2})$. L'écart-type et certains quantiles des distributions empiriques de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ pour les populations gamma, qui sont les plus problématiques. Comme on peut le voir, les distributions de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ sont toujours asymétriques positives et varient considérablement. En d'autres termes, on a besoin d'échantillons plus importants que ceux envisagés ici pour obtenir une estimation fiable de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$. De plus \hat{Y}_{r3} gagnera en efficacité sur \hat{Y}_{r1} quand la valeur réelle de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ dépasse le seuil de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$ pour lequel on passe de \hat{Y}_{r1} à \hat{Y}_{r2} .

6.2 La deuxième étude empirique

Dans la deuxième étude empirique, on envisage une population finie, divisée en huit strates de 100 éléments, connue pour chaque unité de la population. Afin de simuler une stratification articulée sur x , on a attribué les valeurs de x par la fonction monotone de h et de i

$$x_{hi} = 4,95 + 5 \sum_{j=1}^{h-1} j + h \cdot i,$$

où hi indique l'unité $i = 1, 2, \dots, 100$ dans la strate

$$h = 1, 2, \dots, 8.$$

Etant donné x , on a créé une population finie de y valeurs au moyen du modèle

$$Y_{hi} = 20 + 2x_{hi} + 0,06x_{hi}^2 + \epsilon_{hi} \cdot x_{hi},$$

où ϵ_{hi} est une variable aléatoire normale standardiser. La moyenne, l'écart-type et l'indice d'asymétrie de y s'éta-

blissent à 618,2, à 676,0 et à 1,21, respectivement. La corrélation entre y et x s'élève à 0,96.

On s'est servi d'un plan d'échantillonnage aléatoire à stratification proportionnelle pour prélever 5 000 échantillons de taille $n = 40$ (cinq unités par strate) et 2 500 échantillons de 80 éléments (dix par strate). On a calculé les quantités suivantes pour chaque échantillon:

- l'estimateur non biaisé de la moyenne \bar{Y} , pour la population, soit \bar{y} ;
- l'estimateur de ratio \hat{Y}_{r11} , s'appuyant sur le modèle $E_m(Y_{hi}) = \beta x_{hi}$ et $V_m(Y_{hi}) = \sigma^2 x_{hi}$, issu de (5) et de (6) quand $x_{hi} = x_{hi}$ et $v_{hi} = x_{hi}$;
- l'estimateur optimal \hat{Y}_{r21} , reposant sur la même variable auxiliaire que celle utilisée pour \hat{Y}_{r11} ;
- l'estimateur GREG \hat{Y}_{r12} , d'après le modèle $E_m(Y_{hi}) = \alpha + \beta x_{hi}$ et $V_m(Y_{hi}) = \sigma^2 x_{hi}$, tiré de (5) et de (6) quand $x_{hi} = (1, x_{hi})'$ et $v_{hi} = x_{hi}$;
- l'estimateur optimal \hat{Y}_{r22} , selon les mêmes variables auxiliaires que celles utilisées pour \hat{Y}_{r12} ;
- l'estimateur GREG \hat{Y}_{r13} , reposant sur le modèle $E(Y_{hi}) = \alpha + \beta x_{hi} + \gamma x_{hi}^2$ et $V(Y_{hi}) = \sigma^2 x_{hi}^2$ (le modèle réel), et dérivé de (5) et de (6) quand $x_{hi} = (1, x_{hi}, x_{hi}^2)'$ et $v_{hi} = x_{hi}^2$;
- l'estimateur optimal \hat{Y}_{r23} , s'appuyant sur les mêmes variables auxiliaires que \hat{Y}_{r13} ;
- les approximations linéaires $\hat{Y}_{r13}, \hat{Y}_{r12}$ et \hat{Y}_{r23} de $\hat{Y}_{r11}, \hat{Y}_{r12}$ et \hat{Y}_{r22} , respectivement;
- les statistiques $\hat{\lambda}(\hat{Y}_{r1k}, \hat{Y}_{r2k})$, pour $k = 1, 2, 3$;
- les estimateurs assujettis à l'échantillon $\hat{Y}_{r3k}(k = 1, 2, 3)$, qui prennent la valeur \hat{Y}_{r1k} quand $\hat{\lambda}(\hat{Y}_{r1k}, \hat{Y}_{r2k}) \leq 8\%$, et la valeur de \hat{Y}_{r2k} dans les autres cas.

Nous n'avons pas envisagé les estimations de régression séparées, à cause de la petite taille des échantillons dans la state. La population finie est telle que $\lambda(\hat{Y}_{r11}, \hat{Y}_{r21}) = 0,22$, $\lambda(\hat{Y}_{r12}, \hat{Y}_{r22}) = 0,16$, et $\lambda(\hat{Y}_{r13}, \hat{Y}_{r23}) = 0,00$. Précisons qu'en raison du plan d'échantillonnage envisagé, on a $\hat{Y}_{r21} = \hat{Y}_{r22}$, donc on omet \hat{Y}_{r21} .

Le tableau 3 présente les résultats empiriques obtenus par rapport au biais relatif (BR) en pour cent des estimateurs et à l'erreur quadratique moyenne (EQM), après avoir établi que les estimateurs de Horvitz-Thompson sont égaux à 100 dans le dernier cas. Les résultats sont divisés en fonction de la taille de l'échantillon.

Une fois encore, le biais est négligeable. Le pourcentage de réduction de l'EQM réalisable par rapport à la moyenne de l'échantillon augmente avec le nombre de variables auxiliaires. Tel que prévu cependant, \hat{Y}_{r11} et \hat{Y}_{r12} s'avèrent moins efficaces que l'estimateur optimal \hat{Y}_{r22} reposant sur les mêmes variables auxiliaires. La plupart du temps, les statistiques $\hat{\lambda}(\hat{Y}_{r11}, \hat{Y}_{r21})$ et $\hat{\lambda}(\hat{Y}_{r12}, \hat{Y}_{r22})$ ont une valeur supérieure au seuil de 8 %, surtout quand l'échantillon compte 80 éléments. Les estimateurs \hat{Y}_{r31} et \hat{Y}_{r32} , qui dépendent de l'échantillon, sont plus efficaces que \hat{Y}_{r11} et \hat{Y}_{r12} . On le doit à la médiocrité des modèles qui permettent

était de 8x pour les deux strates lors de la première simulation et de 3x lors de la seconde. À la troisième et à la quatrième simulation, on a tiré les valeurs de x d'une variable aléatoire gamma, transformée de façon linéaire, les paramètres ayant été sélectionnés afin de respecter les moyennes et les variances de x et de y pour la strate dans les deux premières simulations, et un indice d'asymétrie de 2,5 pour x (correspondant au ratio entre le troisième moment central et la troisième puissance de l'écart-type). On a ainsi pu étudier les effets d'une forte asymétrie sur les distributions marginales de y et de x .

Les populations ont été bâties pour que $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2}) = 8,1\%$ quand $V(Y|x) = 8x$, et $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2}) = 18,6\%$ quand $V(Y|x) = 3x$. Précisons que le modèle réel utilise l'estimateur de ratio séparé comme estimateur GREG; son usage exigerait néanmoins qu'on connaisse la moyenne de x pour la strate. Or, on la suppose inconnue.

Pour chaque simulation, on a prélevé 10 000 échantillons de 20 éléments (dix par strate) et 5 000 de 40 (vingt par strate). Pour chacun de ces échantillons, on a calculé la valeur de l'estimateur de Horvitz-Thompson $\hat{Y} = \bar{y}$, de $\hat{Y}_{r1}, \hat{Y}_{r2}$, $\hat{Y}_{r1}, \hat{Y}_{r2}$, et de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$. Nous avons aussi calculé un estimateur \hat{Y}_{r3} , qui prend la valeur \hat{Y}_{r1} quand $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2}) \leq 8\%$, et la valeur \hat{Y}_{r2} dans les autres cas. Par conséquent, \hat{Y}_{r3} est un estimateur assujéti à l'échantillon, construit par sélection de \hat{Y}_{r1} ou \hat{Y}_{r2} selon la valeur estimative de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$. Le choix de 8 % est arbitraire. Il s'agit d'un seuil auquel, estime-t-on, il est commode de passer de \hat{Y}_{r1} à \hat{Y}_{r2} .

Tableau 1

Biais relatif (BR) empirique en pour cent et erreur quadratique moyenne (EQM) de \bar{y} , \hat{Y}_{r1} , \hat{Y}_{r2} , \hat{Y}_{r3} et \hat{Y} et proportion de l'échantillon pour laquelle $\hat{\lambda}(\hat{Y}_{r1}, \hat{Y}_{r2}) > 8\%$ dans la première étude empirique

Populations uniformes									
$V(Y x) = 8x$					$V(Y x) = 3x$				
$n = 20$					$n = 20$				
Estimateur	BR (%)	EQM	BR (%)	EQM	Estimateur	BR (%)	EQM	BR (%)	EQM
\bar{y}	-0,06	100,0	-0,08	100,0	\bar{y}	-0,06	100,0	-0,10	100,0
\hat{Y}_{r1}	-0,05	83,8	-0,06	84,1	\hat{Y}_{r1}	-0,05	69,4	-0,05	68,8
\hat{Y}_{r2}	-0,03	77,3	-0,04	77,7	\hat{Y}_{r2}	0,01	56,2	-0,00	55,8
\hat{Y}_{r1}	0,07	87,7	-0,01	86,2	\hat{Y}_{r1}	-0,00	73,4	-0,00	70,5
\hat{Y}_{r2}	-0,05	82,4	-0,04	80,1	\hat{Y}_{r2}	-0,00	59,8	-0,00	57,3
\hat{Y}_{r3}	-0,06	85,0	-0,05	83,1	\hat{Y}_{r3}	-0,01	61,0	-0,01	57,9
Freq ($\lambda > 8\%$)	53,5 %		53,6 %		Freq ($\lambda > 8\%$)	88,6 %		93,5 %	
Populations gamma									
$V(Y x) = 8x$					$V(Y x) = 3x$				
$n = 20$					$n = 20$				
Estimateur	BR (%)	EQM	BR (%)	EQM	Estimateur	BR (%)	EQM	BR (%)	EQM
\bar{y}	0,07	100,0	-0,01	100,0	\bar{y}	-0,03	100,0	-0,03	100,0
\hat{Y}_{r1}	0,08	84,1	0,02	84,3	\hat{Y}_{r1}	-0,03	69,8	-0,03	69,9
\hat{Y}_{r2}	0,09	77,5	0,05	78,1	\hat{Y}_{r2}	-0,02	57,1	-0,02	56,9
\hat{Y}_{r1}	-0,58	88,4	-0,30	86,7	\hat{Y}_{r1}	-0,36	75,5	-0,36	72,8
\hat{Y}_{r2}	0,03	85,8	0,03	80,9	\hat{Y}_{r2}	-0,02	63,5	-0,02	59,1
\hat{Y}_{r3}	-0,05	87,9	0,07	86,2	\hat{Y}_{r3}	-0,04	65,4	-0,04	60,8
Freq ($\lambda > 8\%$)	50,6 %		50,3 %		Freq ($\lambda > 8\%$)	86,9 %		91,7 %	

Le tableau 1 présente les résultats empiriques de chaque simulation par rapport au biais relatif (BR) en pour cent des estimateurs et l'erreur quadratique moyenne (EQM), après avoir établi, dans le second cas, que l'estimateur de Horvitz-Thompson est égal à 100 quand on multiplie l'EQM par 100/EQM(\bar{y}). Comme on peut le constater, le biais est négligeable dans tous les cas (la valeur absolue la plus élevée est inférieure à 0,6 % et le biais est toujours inférieur à 10 % de l'erreur-type correspondante), si bien qu'il ajoute très peu à l'EQM. Le pourcentage de réduction de l'EQM réalisable quand on passe de \hat{Y}_{r1} à \hat{Y}_{r2} correspond à peu près aux valeurs de $\lambda(\hat{Y}_{r1}, \hat{Y}_{r2})$, fixées à l'avance, soit 8,1 % et 18,6 %. Les valeurs réelles de l'EQM de \hat{Y}_{r1} et de \hat{Y}_{r2} dépassent les valeurs asymptotiques correspondantes, surtout quand la population est asymétrique et qu'il s'agit de l'estimateur optimal. Par exemple, dans la troisième simulation, quand $n = 20$, l'EQM de \hat{Y}_{r1} augmente de 5,1 % par rapport à celle de \hat{Y}_{r1} , tandis que la valeur correspondante de \hat{Y}_{r2} est 10,7 %. Lorsqu'on double la taille de l'échantillon, ces valeurs relatives baissent respectivement à 2,8 % et à 3,6 %. Comme nous le soulignons dans l'exemple 2, avec une répartition proportionnelle de l'échantillon, \hat{Y}_{r2} est égal à l'estimateur GREG lorsqu'il s'agit d'un modèle homoscedastique linéaire suivant deux courbes de régression dans les deux strates. La plus grande perte d'efficacité en pour cent de \hat{Y}_{r2} par rapport à sa variance asymptotique trouve donc son explication dans le paramètre supplémentaire que le modèle doit estimer.

$h = 1, 2, \dots, H$. Ce modèle s'ajuste à différentes courbes de régression présentant la même pente, à l'intérieur des strates.

Exemple 3. Voyons un plan d'échantillonnage complexe, et supposons que la population puisse être divisée en H strates a posteriori de taille connue. Soit le modèle de superpopulation $E^m(X'_i) = \beta_{h(i)}, V^m(X'_i) = \sigma^2$, et $C^m(X'_i, X'_j) = 0, i \neq j$, où l'indice $h(i)$ signale la strate a posteriori à laquelle appartient l'unité i . En notant d_{hi} la variable signalant que l'unité i appartient à la strate a posteriori h , et \bar{D}_h correspondant à la moyenne de la population, qu'on connaît, si on établit $x'_i = (d_{i1}, d_{i2}, \dots, d_{iH})'$ et $v'_i = 1$, en (5), on obtient l'estimateur de stratification a posteriori $\bar{X}_{p1} = \sum_{h=1}^H \bar{D}_h \bar{Z}_h / \bar{D}_h$, où \bar{Z}_h et \bar{D}_h sont les estimateurs Horvitz-Thompson de la moyenne des variables $z_{hi} = X'_i d_{hi}$ et \hat{d}_{hi} , respectivement. L'approximation linéaire est $\bar{X}_{p1} = \bar{Y} + (\bar{X} - \bar{Y})\beta_1$, où $\beta_1 = (R_1, R_2, \dots, R_H)'$, $R_h = \bar{Z}_h / \bar{D}_h$ (soit la valeur moyenne de y dans la strate a posteriori h), et $\bar{X} = (\bar{D}_1, \bar{D}_2, \dots, \bar{D}_H)'$. Puisque $U_i = Y_i - \sum_{h=1}^H R_h d_{hi}$, la covariance de \bar{D}_h et \bar{X}_{p1} est

$$C(\bar{X}_{p1}, \bar{D}_h) = C(\bar{Y}, \bar{D}_h) - \sum_{j=1}^J R_j C(\bar{D}_j, \bar{D}_h). \quad (11)$$

En vertu du modèle de superpopulation sur lequel repose \bar{X}_{p1} , $E_{m-1}[C(\bar{X}_{p1}, \bar{D}_h)] = 0$ et on peut s'attendre à ce que $C(\bar{X}_{p1}, \bar{D}_h)$ ait une valeur négligeable pour toutes les valeurs h . On constate aisément que pour un échantillon-nage aléatoire simple, la formule (11) donne zéro. Avec les plans d'échantillonnage complexes cependant, la covariance pourrait prendre une valeur non négligeable, par exemple quand une régression linéaire des totaux des unités primaires de z_{hi} sur les totaux de d_{hi} débouche sur des coordonnées à l'origine non négligeable pour certaines valeurs de h , dans un plan d'échantillonnage à degrés multiples. On trouvera une étude de cas sur la question dans Casady et Valliant (1993).

6. ÉTUDES EMPIRIQUES

L'analyse qui précède s'appuie sur des approximations du premier degré. Les études empiriques qui suivent examinent le fonctionnement des échantillons finis de \bar{Y}_{p1} et de \bar{Y}_{p2} dans le cadre de l'exemple 2.

6.1 La première étude empirique

Dans cette première étude, nous envisageons une population infinie divisée en deux strates de poids égaux et un plan d'échantillonnage aléatoire à stratification proportionnelle, en vue d'estimer la moyenne de la variable d'enquête y . À cette fin, supposons qu'il existe une variable scalaire x dont on ne pouvait se servir pour la stratification, mais dont la moyenne \bar{X} est connue pour la population et

dont la moyenne est inconnue pour la strate (bref, on ignore la valeur de x pour les unités non échantillonnées). Puisque seule la moyenne de x pour la population est connue, une régression linéaire avec erreurs homoscedastiques, c'est-à-dire $E^m(X'_i) = \alpha + x'_i \beta$, $V^m(X'_i) = \sigma^2$, $C^m(X'_i, X'_j) = 0, i \neq j$ peut constituer un modèle de superpopulation raisonnable pour identifier un estimateur GREG. La variable auxiliaire greffée à (5) est $x'_i = (1, x'_i)'$, si bien qu'on peut écrire l'estimateur GREG correspondant de la manière suivante

$$\bar{Y}_{p1} = \bar{y} + (\bar{X} - \bar{x}) s_{yx} / s_x^2,$$

où \bar{y} et \bar{x} représentent les moyennes de y et de x pour l'échantillon, s_{yx} correspond à la covariance entre y et x pour l'échantillon, et s_x^2 est la variance de x pour l'échantillon. L'approximation linéaire est

$$\bar{Y}_{p1} = \bar{y} + (\bar{X} - \bar{x}) S_{yx} / S_x^2,$$

où S_{yx} et S_x^2 sont des analogues de s_{yx} et de s_x^2 pour l'ensemble de la population. Si on laisse tomber le premier élément de $x'_i = (1, x'_i)'$, dont la moyenne estimée n'inclut pas d'erreur, l'estimateur optimal reposant sur la même variable auxiliaire correspond à

$$\bar{Y}_{p2} = \bar{y} + (\bar{X} - \bar{x}) \hat{C}(\bar{y}, \bar{x}) / \hat{V}(\bar{x}),$$

où \bar{X} représente la moyenne de x pour la population, et $\hat{C}(\bar{y}, \bar{x})$ et $\hat{V}(\bar{x})$, les estimateurs non biaisés habituels de la covariance entre \bar{y} et \bar{x} et de la variance de \bar{x} , respectivement. L'approximation linéaire correspondante est

$$\bar{Y}_{p2} = \bar{y} + (\bar{X} - \bar{x}) C(\bar{y}, \bar{x}) / V(\bar{x}),$$

où $C(\bar{y}, \bar{x})$ et $V(\bar{x})$ indiquent la covariance et la variance réelles.

On peut simplifier l'expression $\lambda(\bar{Y}_{p1}, \bar{Y}_{p2})$ afin d'obtenir

$$\lambda(\bar{Y}_{p1}, \bar{Y}_{p2}) = \frac{\sum_1 S_{y2}^2 \sum_2 S_{hx}^2}{\sum_2 S_{yx}^2} \left(\frac{\sum_1 S_{y2}^2}{\sum_2 S_{yx}^2} - \frac{\sum_1 S_{hx}^2}{\sum_2 S_{yx}^2} \right),$$

qu'on peut estimer en remplaçant la variance et la covariance de la population par celles de l'échantillon. On a procédé à quatre simulations. Dans les deux premières, les valeurs de x pour l'échantillon ont été tirées d'une distribution uniforme de [30-70] dans la première strate et de [50-90] dans la deuxième. Sachant x , on a prélevé les valeurs de y pour l'échantillon d'une distribution normale ayant pour valeurs probables 1,26x dans la première strate et 0,82x dans la deuxième. La variance conditionnelle

$$+ \frac{{}^I u^I z N}{{}^I u^I - I} {}^I x^I N \sum_N^{I \neq I} \sum_N^{I=I} = ({}^I \underline{X}^I, {}^I \underline{X}^I) \mathcal{O}$$
$$({}^{1'}\underline{A})A/({}^{1'}\underline{A}, \underline{X})\mathcal{C}_1 - [(\underline{X})A], ({}^{1'}\underline{A}, \underline{X})\mathcal{C} = ({}^{2'}\underline{A}, {}^{1'}\underline{A})\gamma$$
$$6) \quad \left[R - \frac{x}{S} \frac{S}{z} \right] \frac{uN}{u - N} = \left(\frac{1}{\tilde{z}} \frac{\Delta}{x} \right) C$$
$$\hat{\beta}_1 = \frac{\sum_{i \in s} Y_i X_i' / N \pi_i' - \bar{Y} \bar{X}}{\sum_{i \in s} X_i^2 / N \pi_i' - \bar{X}^2},$$

Il est intéressant de noter que si l'échantillon est proportionnellement réparti, soit si $n^h \propto N^h$, en oubliant les termes de grandeur $1/N^h$ par rapport à l'unité, Y^h correspond à l'estimateur GREG qui s'articule sur la variable auxiliaire $x^h = (d_{21}^h, d_{22}^h, \dots, d_{2H}^h, x^h)$ et $v^h = 1$, où d_{2h}^h est une variable indiquant que l'unité h appartient à la strate h .

En prenant $V(\hat{X})$ et $C(\hat{X}, Y)$, on parvient à l'autre estimateur de régression

$$\hat{Y}_{r2} = \hat{Y} + (\bar{X} - \hat{X})' \hat{\beta}_2,$$

où $\hat{\beta}_2 = [V(\hat{X})]^{-1} C(\hat{X}, Y)$. Montanari (1987) l'a étudié et Rao (1994) l'a baptisé l'estimateur optimal. Quand $V(\hat{X})$ est singulier et a pour rang $q' < q$, on doit abandonner une ou plusieurs observations de x_i' , donc de X de façon à obtenir une matrice de variance $q' \times q'$ non singulière qui servira à définir l'estimateur optimal.

Avec $\hat{\beta}_2$, la variance asymptotique de \hat{Y}_{r2} se simplifie en

$$V(\hat{Y}_{r2}) = V(\hat{Y}) - C(\hat{X}, Y)' [V(\hat{X})]^{-1} C(\hat{X}, Y). \quad (7)$$

Les propriétés de l'estimateur optimal sont les suivantes: i) \hat{Y}_{r2} est aussi efficace que \hat{Y}_{r1} , par rapport à l'asymptote, soit $V(\hat{Y}_{r2}) \leq V(\hat{Y}_{r1})$; ii) la moyenne des variables auxiliaires obtenue grâce à l'estimateur optimal est égale à la moyenne connue pour la population correspondante, à savoir $\bar{X}_{r2} = \bar{X}$. Avec l'estimateur GREG, il est possible d'exprimer l'estimateur optimal \hat{Y}_{r2} comme un simple estimateur pondéré quand il y a plus d'une variable d'enquête, en appliquant les mêmes poids à toutes les variables auxquelles on s'intéresse. En prenant la formule de Horvitz-Thompson pour les estimateurs de la variance et de la covariance, on peut écrire $\hat{Y}_{r2} = \sum_{i \in s} Y_i w_i$, où

$$w_i = \frac{1}{N} + (\bar{X} - \hat{X})' [V(\hat{X})]^{-1}$$

$$\left(x_i' \frac{1 - \pi_i}{N^2 \pi_i^2} + \sum_{j \neq i} x_j' \frac{\pi_j - \pi_i \pi_j}{N^2 \pi_i \pi_j \pi_j} \right).$$

La formule de Yates-Grundy parvient à un résultat analogue.

Soulignons que l'optimalité asymptotique de \hat{Y}_{r2} est une propriété qui repose strictement sur le plan d'échantillonnage et qui se réalise à la condition que la population soit finie (bref, dans le contexte de l'approche d'une population fixe pour l'inférence d'une population finie). En revanche, l'optimalité asymptotique de \hat{Y}_{r1} exige que le modèle se vérifie, et elle se rapporte à la variance asymptotique moyenne des populations finies qu'engendre le modèle.

En raison des résultats qui précèdent, \hat{Y}_{r2} paraît préférable à \hat{Y}_{r1} . Cependant, $\hat{\beta}_1$ est une fonction des estimateurs du total de la population, tandis que $\hat{\beta}_2$ est une fonction des estimateurs de la variance et de la covariance. Le premier est donc plus vulnérable à une mauvaise spécification du modèle et le second, aux variations de l'échantillonnage. Avec un échantillon de taille finie, \hat{Y}_{r2} se montre généralement moins stable et plus complexe à calculer, alors que sa variance peut dépasser celle de \hat{Y}_{r1} .

(lire Casady et Valliant 1993). Quoiqu'il en soit, si on dispose d'un nombre de degrés de liberté g suffisant pour estimer β_2 , il est possible de surmonter la difficulté que pose l'instabilité de \hat{Y}_{r2} . Par exemple, avec les plans d'échantillonnage complexes habituels, où on procède par tirage avec remise au premier degré, on peut approximativement considérer g comme le nombre de grappes de l'échantillon moins le nombre de strates (Lehtonen et Pahkinen 1995; p. 181; lire Eltinge et Jang 1996, pour une analyse plus détaillée). On peut s'attendre à ce que β_2 soit stable si g est assez important par rapport à la dimension q de la variable auxiliaire X_1' . Puisque les ordinateurs modernes facilitent de beaucoup le calcul de \hat{Y}_{r2} , établir un critère qui nous aidera à déterminer quand l'emploi d'un tel estimateur devient avantageux présente beaucoup d'intérêt.

5. UN CRITÈRE DE SÉLECTION ENTRE \hat{Y}_{r1} ET \hat{Y}_{r2}

Examinons le théorème suivant:

Théorème: Soit $V(\hat{Y}_r)$ et $V(\hat{Y}_{r2})$, les variances asymptotiques de l'estimateur de régression général \hat{Y}_r et de l'estimateur optimal \hat{Y}_{r2} , respectivement. Dans ce cas,

$$V(\hat{Y}_r) - V(\hat{Y}_{r2}) = C(\hat{X}, Y)' [V(\hat{X})]^{-1} C(\hat{X}, Y). \quad (8)$$

Preuve: En prenant (3) et (7), l'écart entre les variances est

$$V(\hat{Y}_r) - V(\hat{Y}_{r2}) = \hat{\beta}' V(\hat{X}) \hat{\beta} - 2\hat{\beta}' C(\hat{X}, Y) + C(\hat{X}, Y)' [V(\hat{X})]^{-1} C(\hat{X}, Y).$$

Puisque $\hat{\beta}_2 = [V(\hat{X})]^{-1} C(\hat{X}, Y)$ et $\hat{\beta}' C(\hat{X}, Y) = \hat{\beta}' V(\hat{X}) \hat{\beta}_2$ on obtient

$$V(\hat{Y}_r) - V(\hat{Y}_{r2}) = (\hat{\beta} - \hat{\beta}_2)' V(\hat{X}) (\hat{\beta} - \hat{\beta}_2).$$

Mais $C(\hat{X}, Y) = C(\hat{X}, Y) - V(\hat{X}) \hat{\beta} = V(\hat{X}) (\hat{\beta} - \hat{\beta}_2)$ et (8)

s'ensuit.

Précisons que le côté droit de (8) a une forme quadratique positive définie et est égal à zéro si et seulement si $C(\hat{X}, Y) = 0$. Par conséquent, plus la valeur absolue des observations de $C(\hat{X}, Y)$ est faible, plus petit sera l'écart du théorème qui précède est que pour utiliser efficacement la moyenne de la population connue pour toute variable auxiliaire, on doit retenir des estimateurs sans corrélation avec l'estimateur de la moyenne de la variable auxiliaire. Appliquons le théorème à l'estimateur GREG et examinons la k -ième entrée de $C(\hat{X}, Y_{r1})$, qui peut s'écrire comme suit:

$$\hat{Y}_r = \bar{X}'\hat{\beta} + \sum_{i=1}^{\text{les}} \frac{U_i}{N} \pi_i,$$

où $U_i = Y_i - x_i'\hat{\beta}$, il s'ensuit que

$$V(\hat{Y}_r) = \sum_{i=1}^{\text{les}} \sum_{j=1}^{\text{les}} U_i U_j \frac{N^2 \pi_i \pi_j}{\pi_j - \pi_i \pi_j} + \sum_{i=1}^{\text{les}} U_i^2 \frac{N^2 \pi_i}{1 - \pi_i}.$$

La formule de Horvitz-Thompson donne un estimateur à peu près non biaisé de $V(\hat{Y}_r)$

$$\hat{V}(\hat{Y}_r) = \sum_{i=1}^{\text{les}} \hat{U}_i^2 \frac{1}{1 - \pi_i} + \sum_{i=1}^{\text{les}} \sum_{j \neq i}^{\text{les}} \hat{U}_i \hat{U}_j \frac{N^2 \pi_i^2 \pi_j}{N^2 \pi_i \pi_j \pi_j},$$

où $\hat{U}_i = Y_i - x_i'\hat{\beta}$. Quand la taille de l'échantillon est fixe, on peut aussi se servir de l'estimateur de variance Yates-Grundy, à savoir

$$\hat{V}(\hat{Y}_r) = \sum_{i=1}^{\text{les}} \sum_{j \neq i}^{\text{les}} \left(\frac{\pi_j}{N^2 \pi_i \pi_j} - \frac{\pi_i}{N^2 \pi_j} \right) \left(\frac{\hat{U}_i}{\pi_j} - \frac{\hat{U}_j}{\pi_i} \right)^2.$$

Nous appellerons donc $V(\hat{Y}_r)$ la variance asymptotique de \hat{Y}_r .

3. L'ESTIMATEUR DE RÉGRESSION GÉNÉRALISÉ

Deux méthodes peuvent généralement être utilisées pour bâtir le vecteur $\hat{\beta}$. La première a été élaborée dans le cadre de l'approche assistée par modèle servant à tirer des inférences de l'échantillonnage, telle que décrite par Särndal, Swensson et Wretman (1992; sec. 6.4), et Estevao, Hidiroglou et Särndal (1995). Soit Y_i , une variable aléatoire ou son observation, on peut établir le modèle de régression linéaire pour la superpopulation que voici

$$\begin{cases} E^m(Y_i) = x_i' \beta, & i = 1, 2, \dots, N, \\ V^m(Y_i) = \sigma^2 v_i, \\ C^m(Y_i, Y_j) = 0, & i \neq j, \end{cases} \quad (4)$$

où E^m , V^m et C^m indiquent l'espérance mathématique, la variance et la covariance prévues, par rapport au modèle; et β et σ^2 sont des paramètres inconnus du modèle; v_i est une fonction connue de x_i . Le vecteur

$$\bar{\beta}_1 = \left[\sum_{i=1}^N \frac{x_i x_i'}{v_i} \right]^{-1} \sum_{i=1}^N \frac{x_i Y_i}{v_i}$$

est l'estimateur des moindres carrés de β issu du recensement. Si on généralise, comme dans les articles cités en référence,

$$\hat{C}(\hat{X}, \hat{Y}) = \sum_{i \in \text{les}} x_i x_i' \frac{N^2 \pi_i^2}{1 - \pi_i} + \sum_{i \neq j}^{\text{les}} \sum_{j \neq i}^{\text{les}} x_i x_j' \frac{N^2 \pi_i \pi_j}{\pi_j - \pi_i \pi_j}.$$

Examinons maintenant les estimateurs non biaisés $\hat{V}(\hat{X})$ et $\hat{C}(\hat{X}, \hat{Y})$ de $V(\hat{X})$ et $C(\hat{X}, \hat{Y})$, respectivement, dont l'existence dépend de probabilités d'inclusion positives au deuxième degré pour le plan d'échantillonnage. Ces estimateurs dérivent de la formule de Horvitz-Thompson ou de Yates-Grundy, selon le cas. En prenant le premier, par exemple, on obtient le vecteur de la covariance estimée

$$\hat{\beta}_2 = [V(\hat{X})]^{-1} C(\hat{X}, \hat{Y}).$$

Pour trouver un autre estimateur de régression reposant sur la même variable auxiliaire x_i , on recourt à une deuxième approche en vertu de laquelle le vecteur $\hat{\beta}$ minimise la variance asymptotique (3) de l'estimateur de différence (2). En supposant que $V(\hat{X})$ ne présente pas de valeur singulière, c'est-à-dire qu'aucune des composantes de \hat{X} ne se combine linéairement de manière à produire une variance d'échantillonnage nulle, le vecteur de la variance minimale est

4. L'ESTIMATEUR OPTIMAL

Les estimateurs bien connus, dont on se sert couramment, comme l'estimateur de ratio et l'estimateur de stratification a posteriori, appartiennent à la catégorie des estimateurs GREG. Par ailleurs, on a récemment élargi cette catégorie grâce à la technique de calage (Deville et Särndal 1992), afin de mieux contrôler la variabilité des poids finals des observations.

Le modèle est bien spécifié, aucun autre estimateur asymptotique non biaisé n'est plus efficace que \hat{Y}_{r1} , en moyenne (par rapport au modèle). si le modèle est bien spécifié, aucun autre estimateur de plan d'échantillonnage (Wright 1983). En conséquence, parmi les estimateurs asymptotiques non biaisés de \bar{Y} , selon modèle, soit $E^m V(\hat{Y}_{r1})$, donne la valeur la plus faible variance asymptotique de l'échantillonnage que prévoit le de la population, connue, donc $\hat{X}_{r1} = \bar{X}$; ii) la valeur de l'estimateur GREG est égale à la moyenne correspondante de la population, connue, donc $\hat{X}_{r1} = \bar{X}$; i) la moyenne des variables auxiliaires estimées grâce à la partie 2, cet estimateur présente les propriétés suivantes: quand on le remplace dans (1). Outre celles mentionnées à

donne un estimateur convergent de $\hat{\beta}_1$ et l'estimateur de régression généralisé (GREG)

$$\hat{\beta}_1 = \left[\sum_{i=1}^{\text{les}} \frac{x_i x_i'}{\pi_i v_i} \right]^{-1} \sum_{i=1}^{\text{les}} \frac{x_i Y_i}{\pi_i v_i} \quad (5)$$

(6)

Estimation de la moyenne d'une population finie par régression

GIORGIO E. MONTANARI¹

RÉSUMÉ

Dans cet article, l'auteur se penche sur les grandes propriétés de l'estimateur de régression généralisé de la moyenne d'une population finie et de l'estimateur de régression dérivé de la situation et présente un critère qui facilitera le choix plus efficace que le premier, on cerne les conditions à l'origine de la situation et présente un critère qui facilitera le choix entre les deux estimateurs. Une étude de simulation illustre la performance des estimateurs avec un échantillon fini.

MOTS CLÉS: Estimateur de régression généralisé; estimateur de différence; données auxiliaires.

1. INTRODUCTION

L'estimation par régression est une technique efficace pour estimer les moyennes ou les totaux d'une population finie à l'égard des variables d'une enquête, quand on connaît les moyennes ou les totaux de la population pour un jeu de variables auxiliaires. Le problème peut s'énoncer comme suit. Soit une population finie $\mathcal{O} = \{a_1, a_2, \dots, a_N\}$ composée de N unités étiquetées $1, 2, \dots, N$. Soit X_i la valeur de l'unité a_i d'une variable d'enquête y dont la moyenne $\bar{Y} = \sum_{i=1}^N Y_i/N$ doit être estimée à partir d'un échantillon de \mathcal{O} . Supposons pour cela qu'on connaisse la moyenne $\bar{X} = \sum_{i=1}^N x_i/N$ du vecteur d'une variable auxiliaire de dimension q , ayant la valeur $x_i = (x_{i1}, x_{i2}, \dots, x_{iq})'$ pour l'unité a_i , par exemple grâce à un registre administratif ou à un recensement. Les entrées x_i peuvent correspondre à une quantité ou à une variable indicatrice signalant que l'unité appartient à une sous-population précise. Soit s , le jeu d'étiquettes des unités de l'échantillon issu d'un plan d'échantillonnage caractérisé par des probabilités d'inclusion au premier degré $\pi_i, i = 1, 2, \dots, N$, strictement positives. Dans ce cas, l'estimateur de régression peut s'exprimer comme suit:

$$\hat{\bar{Y}}_r = \bar{\bar{Y}} + (\bar{X} - \bar{\bar{X}})' \hat{\beta}_r \quad (1)$$

où $\bar{\bar{Y}} = \sum_{i \in s} Y_i/N\pi_i$ et $\bar{\bar{X}} = \sum_{i \in s} x_i/N\pi_i$ correspondent aux estimateurs non biaisés de Horvitz-Thompson pour \bar{Y} et \bar{X} , respectivement, et $\hat{\beta}_r$ désigne le vecteur des coefficients de régression obtenu par une fonction quelconque des données de l'échantillon $\{(Y_i, x_i'), i \in s\}$. En bref, on calcule $\hat{\bar{Y}}_r$ en ajoutant à l'estimateur non biaisé $\bar{\bar{Y}}$ des termes proportionnels à l'écart entre la moyenne véritable des variables auxiliaires $\bar{X}_k = \sum_{i=1}^N x_{ki}/N, k = 1, 2, \dots, q$ et les estimations correspondantes $\bar{\bar{X}}_k = \sum_{i \in s} x_{ki}/N\pi_i$.

L'article examine les deux principales méthodes servant à construire le vecteur $\hat{\beta}_r$ et les propriétés des estimateurs de régression qui en résultent. On propose ensuite un critère

s'articulant sur une approximation du premier degré permettant de choisir l'une ou l'autre méthode. Enfin, l'auteur donne les résultats de deux études empiriques entreprises en vue d'analyser la performance des deux estimateurs à l'égard des échantillons finis. Les probabilités et les variances sans indice viennent d'un plan d'échantillonnage. L'indice m signale que les calculs se rapportent à un modèle spécifique.

2. PROPRIÉTÉS PRINCIPALES DE L'ESTIMATEUR DE RÉGRESSION

En imposant de légères restrictions aux probabilités d'inclusion du deuxième degré et aux moments qui limitent la population de Y_i et de x_i' , on fait en sorte que l'estimateur $\hat{\bar{Y}}_r$ puisse être calculé approximativement par l'estimateur de différence

$$\hat{\bar{Y}}_r = \bar{\bar{Y}} + (\bar{X} - \bar{\bar{X}})' \hat{\beta}_r \quad (2)$$

où $\hat{\beta}_r$ représente la probabilité limite du vecteur $\hat{\beta}_r$, quand la taille de l'échantillon et de la population tend vers l'infini, les limites correspondant à celles établies par Isaki et Fuller (1992), Wright (1983) et Montanari (1987). On peut étudier la performance de l'estimateur de régression avec les gros échantillons grâce à son approximation linéaire (2). L'estimateur de régression $\hat{\bar{Y}}_r$ n'est à peu près pas biaisé parce que $\hat{\bar{Y}}_r$ ne l'est pas. On peut donc dire que la variance de $\hat{\bar{Y}}_r$ correspond approximativement à celle de $\hat{\bar{Y}}_r$ donnée par

$$V(\hat{\bar{Y}}_r) = V(\bar{\bar{Y}}) + \bar{\bar{Y}}' V(\bar{X}) \bar{\bar{Y}} - 2 \bar{\bar{Y}}' C(\bar{X}, \bar{\bar{Y}}) \quad (3)$$

où $V(\bar{\bar{Y}})$ représente la variance de $\bar{\bar{Y}}$, $V(\bar{X})$ désigne la matrice de la variance à $q \times q$ dimensions de \bar{X} , et $C(\bar{X}, \bar{\bar{Y}})$ est le vecteur à q dimensions de la covariance entre \bar{X} et $\bar{\bar{Y}}$. Puisqu'on peut récrire $\bar{\bar{Y}}_r$ de la façon suivante

¹ Giorgio E. Montanari, Dipartimento di Scienze Statistiche, Università di Perugia, Via A. Pascoli - 06100 Perugia, Italie.

tendance à augmenter avec la réduction du nombre de strates, mais il en ira autant du nombre de degrés de liberté (ce qui réduira $t_{\alpha/2}^*$ ou $t_{\alpha/2}^*$). La réponse à cette question se trouve peut-être au niveau de la population étudiée, si bien que l'expérience acquise lors des enquêtes antérieures peut avoir son utilité.

REMERCIEMENTS

Les auteurs désirent remercier le rédacteur associé et les trois examinateurs pour leurs précieuses suggestions. S. Wang a bénéficié de bourses de l'ASA/NSF/BLS et de la National Security Agency (MDA904-96-1-0029), de la National Science Foundation (DMS-9504589) et du Texas A&M University Scholarly and Creative Activities Program (95-59) pour mener à bien ses recherches. Les opinions émises n'engagent que les auteurs et ne reflètent pas nécessairement les politiques du U.S. Bureau of Labor Statistics.

ANNEXE A

Grâce à l'analyse présentée à la partie 2.2, on sait que $n\hat{p}_A$ a une distribution binomiale $Bin(n, p_A)$. Donc, pour $\hat{p}_A = 0, 1/n, 2/n, \dots, 1$,

$$f(\hat{p}_A | p_A) = \frac{\Gamma(n+1)}{\Gamma(n+2)} \frac{\Gamma(n+2)\Gamma(n\hat{p}_A+1)\Gamma(n(1-\hat{p}_A)+1)}{\Gamma(n+1)} \times$$

$$p_A^{n\hat{p}_A+1} (1-p_A)^{n(1-\hat{p}_A)+1} = k_{\hat{p}_A} (p_A)^{n(1-\hat{p}_A)+1}.$$

Pour chaque valeur (fixe) de \hat{p}_A , la fonction $k_{\hat{p}_A}(p_A)$ donne la densité de probabilité d'une distribution bêta ayant pour paramètres $\omega_1 = n\hat{p}_A + 1$ et $\omega_2 = n(1 - \hat{p}_A) + 1$. Puisque ω_1 et ω_2 auront fort probablement une valeur supérieure à l'unité (au moins dans la plupart des situations réelles), on peut obtenir une approximation raisonnable de $k_{\hat{p}_A}(p_A)$ grâce à la fonction de densité d'une distribution normale de moyenne et de variance équivalentes, soit à peu près \hat{p}_A et $\hat{p}_A(1 - \hat{p}_A)/n$ respectivement.

Si $p_A \sim N(\mu, \sigma^2)$, la distribution a posteriori est

$$h(p_A | \hat{p}_A) = f(\hat{p}_A | p_A) g(p_A) / \int_0^1 f(\hat{p}_A | p_A) g(p_A) dp_A \approx ce^{-\frac{1}{2} \left(\frac{\hat{p}_A(1-\hat{p}_A)n}{(p_A - \hat{p}_A)^2} + \frac{\sigma^2}{(p_A - \mu)^2} \right)}.$$

BIBLIOGRAPHIE

- DORMAN, A., et VALLANT, R. (1993). Quantile variance estimators in complex surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 866-871.
- JOHNSON, E.G., et RUST, K.F. (1993). Effective Degrees of Freedom for Variance Estimates from a Complex Sample Survey. Article présenté à 1993 Joint Statistical Meetings, San Francisco.
- KOTT, P.S. (1994). Test d'hypothèse portant sur des coefficients de régression linéaire et basé sur des données d'enquête. *Techniques d'enquête*, 20, 167-172.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SATTEKTHWAITE, F. (1946). An approximate distribution of estimates of variance components. *Biometrics*, 2, 110-114.

ANNEXE B

Lorsqu'on abandonne l'hypothèse spécifique relative à σ^2 , et qu'on établit $\psi = (\hat{p}_A(1 - \hat{p}_A)/n)/\sigma^2$, il s'ensuit que $[p_A | \hat{p}_A] \sim N(\hat{p}_A, \hat{p}_A(1 - \hat{p}_A)/(1 + \psi)n)$.

Résultat: Supposons que W ait pour distribution $N(0, c^2)$ et qu'à la condition $W = w$, la variable aléatoire T ait une distribution t non centrale avec v degrés de liberté et w pour paramètre de non-centralité. La distribution non conditionnelle de $T/\sqrt{c^2 + 1}$ correspondra à une distribution t centrale avec v degrés de liberté.

Preuve: Soulignons d'abord qu'on peut écrire T sous la forme $T = (X + W)/\sqrt{S^2/v}$, où X a pour distribution $N(0, 1)$, S^2 a pour distribution χ^2_v , et X , W et S^2 sont mutuellement indépendants. Par conséquent, $X' = (X + W)/\sqrt{1 + c^2}$ a pour distribution $N(0, 1)$. Puisque X' et S^2 sont indépendants par définition, $T' = T/\sqrt{1 + c^2} = X'/\sqrt{S^2/v}$ a pour distribution t_v .

- 5) Parmi les autres combinaisons variance-degrés de liberté, c'est $s_{2, sid}^2$ avec v_{max}^2 qui donne le plus petit. Les variations peuvent être appréciables; on doit accepter un compromis entre la couverture et la taille de l'intervalle.
- 6) Pour un type d'intervalle donné, la longueur relative de l'intervalle tend vers 1 lorsque v_{max} augmente. On parvient à des conclusions analogues en étudiant la rémunération moyenne.

Grande population: Le tableau 4 indique la couverture et la grandeur de l'intervalle de confiance de la rémunération globale pour cinq types d'intervalles, classés d'après la valeur moyenne de v_{max} . Ces intervalles comprennent les trois types examinés précédemment avec la petite population, plus deux nouveaux reposant sur le nombre de degrés de liberté pondéré combiné à s_a et à s_b , respectivement. Les résultats sont ceux obtenus après 5 000 répétitions.

- 1) Les résultats sont cohérents avec ceux de la petite population, pour ce qui est de la couverture et de la grandeur relative des différents intervalles. La distribution normale type donne de piètres résultats pour de nombreuses professions.
- 2) L'intervalle reposant sur le nombre de degrés de liberté pondéré v_1 , ne donne une couverture de moins de 90 % que pour un petit nombre de cas.
- 3) La grandeur de l'intervalle peut varier fortement avec le genre d'intervalle. Néanmoins, à mesure que v_{max} augmente, le ratio de la grandeur de l'intervalle avec le quadruple de l'erreur quadratique moyenne a tendance à approcher la valeur 1.
- 4) Utiliser s_a, s_b , ou s_{sid} avec t_{q_1} ne présente guère de différence.

En dépit de distinctions mineures, les résultats obtenus avec la grande population nous amènent aux mêmes deductions, que nous ne répéterons pas ici.

4. RÉSUMÉ ET CONCLUSIONS

De notre analyse théorique et de l'étude de simulation, on peut tirer les conclusions qui suivent.

1. Les intervalles de confiance type de 95 % pour la moyenne ou le total d'un domaine ont tendance à assurer une couverture inférieure à 95 % quand ils reposent sur une distribution normale type et les méthodes types d'estimation de la variance. L'écart varie avec le domaine (la profession, dans l'étude sur la rémunération), mais peut être assez important, même avec un gros échantillon.

2. Les nouvelles méthodes améliorent nettement la situation puisque les intervalles assurent une plus grande couverture, typiquement égale à 95 % ou presque. Les nouveaux intervalles ont tendance à être plus grands que les intervalles type. L'élargissement variera avec le domaine et dépendra de la méthode particulière de construction des IC. Les nouveaux intervalles différeront peu des intervalles type quand le domaine engendre des échantillons très importants.
3. Les cas où la couverture tombe sous la couverture nominale, même après rajustement de l'intervalle par t_1 , résultent apparemment d'une sérieuse entorse à l'hypothèse d'une distribution normale des données du domaine. La correction t ne constitue pas une panacée. Quoiqu'il en soit, on note une nette amélioration de la couverture par rapport à celle de l'intervalle type reposant sur la distribution normale.
4. Le but, en créant les nouveaux intervalles, consiste à conditionner par rapport à la somme d'information relative à une profession, grossièrement jugée d'après le nombre d'unités échantillonnées appartenant au domaine concerné. On ignore la proportion d'unités qui se retrouve dans chaque strate. Pour y remédier, on attribue une distribution a priori à l'inconnue, distribution reposant sur ce que nous savons au sujet de l'inconnue, une idée empruntée aux bayésiens. En dernière analyse cependant, c'est la probabilité de couverture réelle qui détermine la valeur d'une telle approche.
5. Ces modifications ont pour principale conséquence l'abandon des quantiles type de la distribution normale ($\pm 1,96$ pour une couverture de 95 %) pour la construction des IC. On les remplace par les quantiles de la distribution t de Student, le nombre de degrés de liberté étant déterminé par l'échantillon et variant avec le domaine. S'il faut indiquer l'écart-type au lieu des intervalles de confiance, à cause des exigences de publication ou pour d'autres raisons, on devrait indiquer un écart-type réel correspondant à la grandeur de l'intervalle de confiance de 95 % articulé sur t et divisé deux fois par 1,96.
6. L'estimation type de la variance paraît tenir quand elle accompagne le nouveau quantile t . Dans la plupart des cas, pareille combinaison devrait s'avérer fort satisfaisante, si bien que la seule différence avec la méthode habituelle concernera le nombre de degrés de liberté ajusté. Il arrive néanmoins que les autres écarts-types améliorent la couverture ou raccourcissent les intervalles de confiance.
7. La mesure et la manière dont les strates devraient être regroupées (s'il y a lieu) en vue d'une estimation de la variance et du nombre de degrés de liberté au moment de construire les intervalles de confiance est une question qui demeure sans réponse. En général, on devra accepter un compromis; la variance aura

Petite population: Le tableau 3 indique la couverture et la grandeur médiane relative de l'intervalle de confiance pour la rémunération globale des échantillons $n_k = 4$ et $n_k = 8$, à l'égard de 8 professions et de trois méthodes de construction différentes des intervalles de confiance: l'estimateur type de la variance s_{sid}^2 , avec quantile z type à distribution normale, le nombre de degrés de liberté non pondéré v_{max} , le nombre de degrés de liberté pondéré v_1 . Les professions sont classées par ordre croissant, selon le nombre moyen de degrés de liberté non pondéré, pour l'ensemble des répétitions. On peut formuler les remarques suivantes:

- 1) L'estimateur type de la variance et les quantiles type à distribution normale (nombre infini de degrés de liberté) donnent presque toujours une piètre couverture. Les autres sortes d'intervalles de confiance aboutissent à des résultats nettement plus satisfaisants. En général, la couverture approche la valeur nominale de 95 % ou est légèrement conservatrice avec les degrés de liberté pondérés; comme prévu, lorsqu'ils reposent sur des degrés de liberté non pondérés cependant, les intervalles de confiance ont tendance à assurer une couverture de quelques points inférieure à celle obtenue avec des degrés de liberté pondérés.
- 2) Deux professions (1122 et 4021) présentent une couverture excessivement faible pour les totaux, même avec les meilleures méthodes proposées. L'étude de ces professions donne à penser que l'hypothèse de la normalité se trouve sérieusement compromise. Ainsi, pour la profession 4021, deux unités de la cinquième strate se caractérisent par une population de travailleurs, donc une rémunération globale, d'un ordre de grandeur supérieur à celle des autres établissements de la strate, et à la population dans son ensemble. D'autre part, le taux de rémunération de ces deux établissements aberrants se situe nettement sous la vaste majorité des établissements. En effet, si on les exclut de la population, la rémunération moyenne s'élève à 9,68 \$/h tandis que lorsqu'on les inclut, elle se situe à 8,28 \$/h. Puisque la cinquième strate comporte 66 établissements, ces deux établissements peuvent aisément ne pas se retrouver dans un échantillon de 8 éléments. Il s'ensuit une grave surestimation de la rémunération moyenne ou une sous-estimation de la rémunération globale. Parallèlement, la rémunération est assez homogène parmi les établissements, de sorte que l'estimation de la variance aura tendance à être trop faible. La présence de plusieurs petits établissements concourt à accroître le nombre de degrés de liberté, et la correction t ne peut entièrement compenser la différence. Il est difficile d'établir comment éviter le problème si ce n'est grâce à des informations préalables, en affectant les établissements aberrants à une strate précise. Malgré cela, les nouveaux intervalles marquent une amélioration sensible sur les intervalles dérivant d'une distribution naïvement normale.

L'étude se bornait aussi à une couverture de 95 %.

3.3 Étude empirique de l'échantillonnage aléatoire stratifié: données sur la rémunération du BLS

Après de mieux estimer la précision des données sur la rémunération fournies par le U.S. Bureau of Labor Statistics, nous avons examiné l'importance de la couverture et des intervalles de confiance dans le cadre de deux études de simulation portant sur des populations tirées d'un échantillon expérimental de l'Occupational Compensation Survey Program (OCCSP) de 1991. L'OCCSP est une enquête menée auprès des établissements de plusieurs régions métropolitaines pour estimer le taux de rémunération dans certaines professions. L'enquête reposait sur un échantillonnage aléatoire simple, les établissements étant stratifiés selon l'importance de leurs effectifs et le secteur d'activité.

Une population (la «petite») correspond à la population de l'échantillon expérimental proprement dit. Elle comprenait six strates incertaines et une certaine de 12 établissements. Cinq cents échantillons stratifiés de taille $n = 36$ et 60, ont été prélevés au hasard à partir de cette population, ce qui représentait les choix $n_k = 4$ et $n_k = 8$, reflétant la taille relative des échantillons tirés de la population originale. On a obtenu la deuxième population (la «grande») en étendant les données de l'échantillon par répétition (échantillonnage aléatoire simple par tirage avec remise dans chaque strate de la petite population) jusqu'à parvenir à une population de la même taille que l'originale. Encore une fois, on comptait six strates incertaines et une certaine. La taille de l'échantillon de chaque strate était identique à celle de l'échantillon réel. Les domaines sont les professions auxquelles on s'intéresse. Seule une fraction des établissements compte des travailleurs d'une profession particulière et se retrouvent dans le domaine correspondant. Le tableau 2 indique le nombre d'établissements ayant des employés dans les professions retenues au sein de la petite population.

Tableau 2
Nombre d'établissements par domaine (profession) et par strate, petite population

strate													
Profession		1	2	3	4	5	6	7	total	4021	1141	1122	3180
		0	4	11	10	8	10	7	50	56	48	56	80
		0	3	11	7	11	9	7	48	56	48	56	80
		0	3	8	13	14	12	6	56	56	48	56	80
		10	11	5	25	20	4	5	80	56	48	56	80
		0	3	14	2	13	17	7	56	56	48	56	80
		2	8	15	9	15	19	9	77	56	48	56	80
		17	20	5	61	31	3	1	138	56	48	56	80
		12	16	22	28	25	27	9	139	56	48	56	80
		35	35	33	136	66	36	12	353	56	48	56	80
Tout établi.		35	35	33	136	66	36	12	353	56	48	56	80

L'échantillonnage s'est effectué par tirage sans remise dans les deux cas, en sorte que les IC incluent des facteurs de correction pour une population finie (le cas échéant). L'étude se bornait aussi à une couverture de 95 %.

considérerons spécifiquement v_{\max} comme le nombre de degrés de liberté non pondéré. Dans ce cas, la limite supérieure de l'IC serait

$$u = \hat{I}_A + \frac{\sqrt{\sum_{k \in B_2} d_k(n_{A_k} - 1)(\hat{\sigma}_A^2 / \sigma_A^2)}}{\Theta^{1/2} t_{v_{\max}}},$$

On pourrait aussi essayer de contourner le problème que pose l'estimation de Θ (du moins quand $B_1 = B_2$) en choisissant judicieusement d_k . Pour cela, supposons que

$$B_1 = B_2 \text{ et soit}$$

$$d_k = \frac{N_k^2 \hat{p}_{A_k} \sigma_A^2}{n_k(n_{A_k} - 1)} \frac{(\gamma_{A_k}^2(1 - \hat{p}_{A_k}) + 1)}{(\gamma_{A_k}^2(1 - \hat{p}_{A_k}) + 1)}$$

si bien que $\sum_{k \in B_2} d_k(n_{A_k} - 1) = \Theta$ et que les termes Θ s'annulent dans (10). On obtient ainsi

$$u = \hat{I}_A + \sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{A_k} \sigma_A^2}{n_k} \frac{(\gamma_{A_k}^2(1 - \hat{p}_{A_k}) + 1)}{(\gamma_{A_k}^2(1 - \hat{p}_{A_k}) + 1)} t_{v_1}},$$

où v_1 représente le nombre de degrés de liberté associé au deuxième d_k choisi. De manière plus générale (à savoir, quand $B_1 \neq B_2$), on obtient

$$u = \hat{I}_A + \frac{\sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{A_k} \sigma_A^2}{n_k} \frac{(\gamma_{A_k}^2(1 - \hat{p}_{A_k}) + 1)}{(\gamma_{A_k}^2(1 - \hat{p}_{A_k}) + 1)}}}{\Theta^{1/2} t_{v_1}},$$

Le problème de l'estimation des paramètres de la population demeure, mais s'y ajoute celui de l'estimation du nombre de degrés de liberté.

Une troisième possibilité, déjà mentionnée, consisterait à établir que $d_k = N_k^2 \hat{p}_{A_k} \sigma_A^2 / n_k(n_{A_k} - 1)$, de sorte que lorsque $B_1 = B_2$, $\hat{\sigma}_A^2 = \hat{\sigma}_A^2 \equiv \sum_{k \in B_2} d_k(n_{A_k} - 1) \hat{\sigma}_A^2 / \sigma_A^2$ devient un estimateur non biaisé de $\hat{\sigma}_A^2$. Nous avons alors

$$u = \hat{I}_A + \frac{\sqrt{\sum_{k \in B_2} \frac{N_k^2 \hat{p}_{A_k} \sigma_A^2}{n_k}}}{\Theta^{1/2} t_{v_2}},$$

où v_2 donne le nombre de degrés de liberté associé au troisième terme d_k . Comme cela se produit dans le deuxième cas, il faut estimer les paramètres de la population et le nombre de degrés de liberté.

Il conviendrait maintenant de noter que si on estime σ_A^2 avec $\hat{\sigma}_A^2$ pour $k \in B_2$ et si Θ est un estimateur à spécifier ultérieurement, les limites supérieures (estimées) ci-dessus

deviennent respectivement $u = \hat{I}_A + \Theta^{1/2} t_{v_{\max}}$, $u = \hat{I}_A + \Theta^{1/2} t_{v_1}$ et $u = \hat{I}_A + \Theta^{1/2} t_{v_2}$. On estime le nombre de degrés de liberté en remplaçant les paramètres estimés de la population se rapportant aux deux valeurs retenues pour d_k . v_1 et v_2 sont plus petits que v_{\max} , si bien que pour toute valeur réelle de Θ , c'est l'intervalle de confiance bâti sur v_{\max} qui sera le plus court. Il n'existe aucun lien général entre la taille de v_1 et de v_2 . Les preuves empiriques indiquent que la deuxième et la troisième approche ne présentent pas une grande différence.

Pour estimer Θ , écrivons

$$\Theta = \sum_{k \in B_1 - B_2} N_k^2 \hat{p}_{A_k} (\mu_{A_k}^2(1 - \hat{p}_{A_k}) + \sigma_A^2) / n_k + \sum_{k \in B_2} N_k^2 \hat{p}_{A_k} (\mu_{A_k}^2(1 - \hat{p}_{A_k}) + \sigma_A^2) / n_k.$$

Pour $k \in B_1 - B_2$, l'estimateur $\hat{\sigma}_A^2$ n'est pas défini, mais il est assez simple de vérifier que $(1 - \hat{p}_{A_k})E[\mu_{A_k}^2 | n_{A_k}] \leq \sigma_{A_k}^2 + \mu_{A_k}^2(1 - \hat{p}_{A_k}) \leq E[\mu_{A_k}^2 | n_{A_k}]$. Il s'ensuit que

$$s_a^2 = \sum_{k \in B_1} N_k^2 \hat{p}_{A_k} (1 - \hat{p}_{A_k}) \mu_{A_k}^2 / n_k + \sum_{k \in B_2} N_k^2 \hat{p}_{A_k} \hat{\sigma}_A^2 (1 + 1/n_k - 1/n_{A_k}) / n_k$$

aura tendance à sous-estimer Θ , et

$$s_b^2 = \sum_{k \in B_1 - B_2} N_k^2 \hat{p}_{A_k} \mu_{A_k}^2 / n_k + \sum_{k \in B_1} N_k^2 \hat{p}_{A_k} (1 - \hat{p}_{A_k}) \mu_{A_k}^2 / n_k + \sum_{k \in B_2} N_k^2 \hat{p}_{A_k} \hat{\sigma}_A^2 (1 + 1/n_k - 1/n_{A_k}) / n_k$$

fera l'inverse. De toute évidence, $s_a^2 \leq s_b^2$ et on ne parvient à l'égalité que quand $B_1 = B_2$.

On peut aussi s'assurer qu'avec l'échantillonnage stratifié, l'estimateur type de la variance des totaux estimés de la population est

$$s_2^{\text{sid}} = \sum_{k \in B_1} N_k^2 s_k^2 / n_k = \sum_{k \in B_1} N_k^2 \hat{p}_{A_k} (1 - \hat{p}_{A_k}) \mu_{A_k}^2 / (n_k - 1) + \sum_{k \in B_2} N_k^2 \hat{p}_{A_k} \hat{\sigma}_A^2 (1 - 1/n_{A_k}) / (n_k - 1).$$

Cet estimateur de Θ est satisfaisant quand n_k n'a pas une faible valeur.

Les résultats qui précèdent impliquent que les IC de la forme $(\hat{I}_A \pm s_b t_{1-a/2, v_1}^{\text{sid}})$ assureront la meilleure couverture; ceux ressemblant à $(\hat{I}_A \pm s_{\text{sid}} t_{1-a/2, v_{\max}}^{\text{sid}})$, voire à $(\hat{I}_A \pm s_{\text{sid}} t_{1-a/2, v_1}^{\text{sid}})$ présentent toutefois des avantages évidents sur le plan des calculs. Plusieurs de ces formes concurrentes sont évaluées empiriquement à la partie 3.3. On peut en étendre aisément les résultats aux estimateurs de ratio par la méthode classique de linéarisation.

où $\hat{p}_A = [\hat{p}_{A1} \hat{p}_{A2} \dots \hat{p}_{AK}]$, $p_A = [p_{A1} p_{A2} \dots p_{AK}]$. Par conséquent, comme pour l'échantillonnage aléatoire simple, il existe un biais conditionnel $\hat{\mu}_A$, dont on doit tenir compte.

3.2 Méthode pour les intervalles de confiance

La méthode générale élaborée à la partie 2.3 pour l'échantillonnage aléatoire simple s'applique également ici. Il suffit de donner aux valeurs scalaires une forme vectorielle, par exemple, remplacer \hat{p}_A par $\hat{p}_A = (\hat{p}_{A1}, \dots, \hat{p}_{AK})'$. Plus précisément $H(x|\hat{p}_A, p_A) = \Pr\{\theta \leq x|\hat{p}_A, p_A\}$ servira de fonction de distribution conditionnelle à $\theta = (T_A - T_A)/\hat{\sigma}_A$, où $\hat{\sigma}_A$ désigne un facteur de remise à l'échelle qui reste à déterminer.

Soit $B_2 = \{k | n_{qk}^A \geq 2 \text{ and } 1 \leq k \leq K\}$ et, pour $k \in B_2$, définissons $\hat{\sigma}_{Ak}^2 = \sum_{i=1}^{n_{qk}^A} (x_{ki} - \hat{\mu}_{Ak})^2 / (n_{qk}^A - 1)$. Sous l'hypothèse de normalité, $(n_{qk}^A - 1)\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2 \sim \chi^2(n_{qk}^A - 1)$. Si $\{d_k | k \in B_2\}$ sont des constantes non négatives pour lesquelles $\sum_{k \in B_2} d_k^2 > 0$, avec l'approximation habituelle à deux moments de Satterthwaite (1946), la variable aléatoire conditionnelle

$$\left[(1/c) \sum_{k \in B_2} d_k^2 (n_{qk}^A - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2) | \hat{p}_A, p_A \right]$$

à peu près pour distribution $\chi^2(v)$, où

$$c = \sum_{k \in B_2} d_k^2 (n_{qk}^A - 1) / \sum_{k \in B_2} d_k^2 (n_{qk}^A - 1)$$

et

$$v = \left(\sum_{k \in B_2} d_k^2 (n_{qk}^A - 1) \right)^2 / \sum_{k \in B_2} d_k^2 (n_{qk}^A - 1).$$

On en déduit qu'on devrait se limiter aux expressions de la

forme générale

$$\hat{\sigma}_A^2 = \sum_{k \in B_2} d_k^2 (n_{qk}^A - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$$

en choisissant le terme d_k à spécifier. Soulignons que lorsque $B_1 = B_2$ et $d_k = N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 / n_k (n_{qk}^A - 1)$, $\hat{\sigma}_A^2 = \hat{\sigma}_{A1}^2 = \sum_{k \in B_2} d_k^2 (n_{qk}^A - 1) \hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2$ est un estimateur non biaisé de la variance conditionnelle σ_A^2 . Néanmoins, comme cela se produit avec l'échantillonnage aléatoire simple, cet

estimateur aura tendance à être trop faible. Nous nous servons donc de l'expression plus générale pour créer une série de statistiques t au moment d'enlever les conditions applicables à p_A . Chacune de ces statistiques fera intervenir des paramètres inconnus, qu'il faudra estimer, ainsi qu'on l'avait fait pour l'échantillonnage aléatoire simple (passage de l'équation (6) à (7)). De ce travail il résultera plusieurs statistiques rivales presque similaires à la statistique t , qu'on pourra ensuite comparer de façon empirique.

Comme les échantillons sont sélectionnés indépendamment de chaque strate, on obtient $f(\hat{p}_A | p_A) = \prod_{k=1}^K f_k(\hat{p}_{Ak} | p_{Ak})$ et, à cause du plan d'échantillonnage des strates, $n_k \hat{p}_{Ak}$ se caractérise par une distribution binomiale

$B(n_k, p_{Ak})$. On présume que les termes $\{p_{Ak} | 1 \leq k \leq K\}$ sont conjointement indépendants et que $g(p_A) = \prod_{k=1}^K g_k(p_{Ak})$, ce qui signifie que

$$f(\hat{p}_A | p_A) g(p_A) = \prod_{k=1}^K f_k(\hat{p}_{Ak} | p_{Ak}) g_k(p_{Ak})$$

et

$$h(\hat{p}_A) = \prod_{k=1}^K \int f_k(\hat{p}_{Ak} | p_{Ak}) g_k(p_{Ak}) dp_{Ak}.$$

Dans ce qui suit, nous présumons que la distribution a priori de p_{Ak} est $N(\mu_{p_{Ak}}, \sigma_{p_{Ak}}^2)$. Pour l'approche empirique de Bayes, on se servira de $\mu_{p_{Ak}} = \hat{p}_{p_{Ak}}$. Comme pour l'échantillonnage aléatoire simple, définissons

$$\psi_{Ak} = \hat{p}_{Ak} (1 - \hat{p}_{Ak}) / n_k \sigma_{p_{Ak}}^2.$$

Il est facile d'étendre le résultat de l'annexe A à l'échantillonnage aléatoire stratifié, d'où il découle que, pour $\hat{\mu}_A$ défini par (9), $[\hat{\mu}_A / \hat{\sigma}_A | \hat{p}_A]$ a pour distribution $N(0, \text{var}(\hat{\mu}_A | \hat{p}_A) / \hat{\sigma}_A^2)$, où $\text{var}(\hat{\mu}_A | \hat{p}_A) = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} \hat{p}_{Ak} (1 - \hat{p}_{Ak}) / n_k (1 + \psi_{Ak})$. En prenant les résultats de l'annexe B, il résulte que, sous réserve de \hat{p}_A , la variable aléatoire

$$\hat{\theta} = \frac{(T_A - T_A) / \sqrt{\text{var}(\hat{\mu}_A | \hat{p}_A) + \hat{\sigma}_A^2}}{\sqrt{\hat{\sigma}_A^2 / cv}} = \frac{(T_A - T_A) / \sqrt{\text{var}(\hat{\mu}_A | \hat{p}_A) + \hat{\sigma}_A^2}}{\sqrt{\hat{\sigma}_A^2 / cv}}$$

$$\frac{\sqrt{\sum_{k \in B_1} d_k^2 (n_{qk}^A - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2)} / \sum_{k \in B_1} d_k^2 (n_{qk}^A - 1)}{(T_A - T_A) / \sqrt{\text{var}(\hat{\mu}_A | \hat{p}_A) + \hat{\sigma}_A^2}}$$

à peu près une distribution t centrale à v degrés de liberté. Soit $\Theta = \text{var}(\hat{\mu}_A | \hat{p}_A) + \hat{\sigma}_A^2$, où

$$\gamma_{Ak}^2 = \mu_{Ak}^2 / \sigma_{Ak}^2$$

en supposant que ψ_{Ak} approche zéro, on a

$$\Theta = \sum_{k \in B_1} N_k^2 \hat{p}_{Ak} \sigma_{Ak}^2 (\gamma_{Ak}^2 (1 - \hat{p}_{Ak}) + 1).$$

La limite supérieure de l'IC serait donc (approxima-

tivement)

$$n = T_A + \frac{\sqrt{\sum_{k \in B_2} d_k^2 (n_{qk}^A - 1) (\hat{\sigma}_{Ak}^2 / \sigma_{Ak}^2)}}{\sqrt{\sum_{k \in B_2} d_k^2 (n_{qk}^A - 1)}} \Theta^{1/2} t_v, \quad (10)$$

où t_v représente les valeurs critiques de la distribution t_v . Malheureusement, les limites ne dépendent pas seulement du terme d_k , choisis, mais aussi des paramètres inconnus μ_{Ak} et σ_{Ak}^2 .

Il est facile de voir que $v \leq \sum_{k \in B_1} (n_{qk}^A - 1) \equiv v_{\max}$ et, si on établit $d_k^* = 1$ (ou toute autre constante), $v = v_{\max}$. Nous

cette étude.

r sultats.

conservatrice, peu importe la valeur de γ_A .

est important.

10

$$\sum_{k \in B_1} N_2^k \hat{p}_2^{Ak} \sigma_2^{Ak} / n^k \equiv \tilde{\sigma}_2^{Ak},$$

$$(6) \quad \tilde{\mathbf{u}} \equiv \mathbf{u}(\mathbf{d}^k - \hat{\mathbf{d}}^k) \mathbf{N} \sum_{l=1}^K = \left[\mathbf{d}^k, \hat{\mathbf{d}}^k \right] (T_L - T) E$$

ment s'assurer que

on $\hat{p}^{A_k} = n^{A_k}/n^k$, $\hat{q}^{A_k} = \sum_{i=1}^{n^k} x_{k,i}/n^{A_k}$ et $B_1 = \{k | n^{A_k} \geq 1 \text{ et } 1 \leq k \leq K\}$. Puisque $\hat{p}^{A_k} = 0$ pour $k \in B_1$, on peut facile-

$$\hat{T}_A = \sum_{k \in B_1} \hat{T}_{Ak} = \sum_{k \in B_1} N_k \hat{p}_{Ak} \hat{q}_{Ak},$$

$T_A = \sum_{k=1}^K x_{ki} \sum_{i \in A} x_{ki} = \sum_{k=1}^K N_k \hat{p}_{A^k} \mu_{A^k}$ serait donc

Supposons qu'il existe K strates et que les termes appropriés, déjà définis, aient leur équivalent dans chaque strate. Par exemple, n_k indiquerait la taille de l'échantillon et n_k^A , le nombre d'éléments échantillonnés dans A à la k -ième strate. Un estimateur naturel du total du domaine

3.1 Définitions et notation

STRATIFIE

3. ÉCHANTILLONNAGE ALÉATOIRE

* Voir l'équation (8) et le texte l'accompagnant pour la définition de IC_1 et IC_3 . IC_0 correspond à l'intervalle de confiance normal type.

P_A	n	M	IC_0	IC_1	IC_3
0,01	100	38774	100,0	100,0	91,2
	300	11773	98,3	100,0	83,2
0,02	100	16327	91,1	99,4	95,0
	300	10078	88,6	95,5	93,9
0,05	100	10303	88,7	97,8	93,5
	300	10000	92,3	94,4	92,5
0,10	100	10001	90,9	94,8	92,5
	300	10000	94,0	95,0	92,3
$y = 1/2$					
0,01	100	37749	99,9	100,0	83,5
	300	11740	94,4	100,0	89,1
0,02	100	16348	99,0	100,0	88,4
	300	10075	91,4	98,9	74,7
0,05	100	10312	90,5	99,5	77,6
	300	10000	93,8	95,8	66,6
0,10	100	10000	91,7	96,5	67,9
	300	10000	94,0	95,2	65,0

Tableau I

Couverture du total du domaine par les intervalles de confiance à 95 %

dans les populations artificielles où les variables du domaine sont normalement distribuées*

Coverture

a) $[\sqrt{n}(\hat{T}_A - T_A)/n | \hat{P}_A, P_A]$ a pour distribution $N(\sqrt{n}\mu_A(\hat{P}_A - P_A), \hat{P}_A\sigma_A^2)$

$$\left[\frac{\sigma_A^2}{n} (\hat{P}_A - 1) \mid \hat{P}_A, P_A \right] \text{ a pour distribution } \chi^2(n\hat{P}_A - 1), \text{ et}$$

c) la variable aléatoire conditionnelle de b) est stochastiquement indépendante de la variable aléatoire conditionnelle de a).

Envisageons $\hat{\theta}_1 = (\hat{T}_A - T_A)/(\hat{N}\hat{\sigma}_A\sqrt{\hat{P}_A}/\sqrt{n})$, qui prend

la variance conditionnelle de \hat{T}_A pour la normalisation. De a), de b) et de c), il s'ensuit que, sous réserve de (\hat{P}_A, P_A) , la variable aléatoire $\hat{\theta}_1$ a une distribution t non-centrale avec $n\hat{P}_A - 1 = n_A - 1$ degrés de liberté et un paramètre de non-centralité

$$\lambda = \sqrt{n}\gamma_A(\hat{P}_A - P_A)/\sqrt{\hat{P}_A},$$

où

$$\gamma_A = \mu_A/\sigma_A.$$

On a donc spécifié la fonction de distribution conditionnelle $H(\cdot | \hat{P}_A, P_A)$ de $\hat{\theta}_1$. Puisque $f(\hat{P}_A | P_A)$ et $g(P_A)$ ont déjà été spécifiées, $F(\cdot | \hat{P}_A)$ en (4) est bien défini, même si la fonction exige des calculs considérablement laborieux. Il convient de souligner qu'il y a dépendance à μ_A et à σ_A^2 , par le truchement de γ_A .

Quoique $F(\cdot | \hat{P}_A)$, tel que mentionné ci-dessus, puisse servir à établir les valeurs critiques, ces dernières s'avèrent très difficiles à calculer. Une approche relativement simple, présentée au paragraphe suivant, donne une bonne approximation de ces valeurs. Nous avons vérifié la précision des approximations en calculant la valeur exacte de certains cas, à partir de simulations à grande échelle.

L'adoption d'une distribution a priori articulée sur P_A débouche sur la distribution a posteriori approximative $P_A \sim N(\hat{P}_A, \text{var}(\hat{P}_A))$ et permet de calculer la valeur approximative de $\text{var}(\hat{P}_A)$ par $\hat{P}_A(1 - \hat{P}_A)/n$. Nous avons préféré la distribution a priori légèrement plus souple $P_A \sim N(\mu, \sigma_A^2)$, et choisi empiriquement $\mu = \hat{P}_A$, en examinant plusieurs possibilités pour σ_A^2 , que nous spécifierons plus loin. De l'annexe A il découle que $[\lambda | \hat{P}_A]$ a une distribution à peu près normale, avec une moyenne de zéro et une variance de $\gamma_A^2(1 - \hat{P}_A)/(1 + \psi_A)$, où

$$\psi_A = \hat{P}_A(1 - \hat{P}_A)/n\sigma_A^2.$$

Partant du résultat de l'annexe B, sous réserve de \hat{P}_A ,

$$\frac{N\hat{\sigma}_A\sqrt{\hat{P}_A}}{(\hat{T}_A - T_A)} \sqrt{\frac{\gamma_A^2(1 - \hat{P}_A)}{1 + \psi_A}} + 1$$

a une distribution t centrale avec $n_A - 1$ degrés de liberté. Soit $t_{1-\alpha/2, n_A-1}$, le $(1 - \alpha/2)$ 100 % percentile de la distri-

$$n = \hat{T}_A + N\hat{\sigma}_A\sqrt{\hat{P}_A}/n \times \left(\left(\gamma_A^2(1 - \hat{P}_A) + 1 + \psi_A \right) / \left(1 + \psi_A \right) \right)^{1/2} t_{1-\alpha/2, n_A-1} \quad (6)$$

Puisque $\hat{\sigma}_A^2$ ne présente conditionnellement pas de biais pour σ_A^2 et puisque $\hat{\mu}_A^2 - \hat{\sigma}_A^2/n_A$ n'en présente conditionnellement pas non plus pour μ_A^2 , on se sert de $\hat{\gamma}_A^2 = (\hat{\mu}_A^2 - \hat{\sigma}_A^2/n_A)/\hat{\sigma}_A^2$ pour estimer γ_A^2 . Si on remplace γ_A^2 par $\hat{\gamma}_A^2$ dans (6), on arrive à

$$\hat{n} \approx \hat{T}_A + (N\hat{\sigma}_A/\sqrt{n}) \times \left(\left(1 + \frac{\hat{P}_A\hat{\sigma}_A^2\hat{\psi}_A}{s^2} \right) / \left(1 + \psi_A \right) \right)^{1/2} t_{1-\alpha/2, n_A-1} \quad (7)$$

où s^2 est défini en (1).

Il ne reste qu'à choisir ψ_A . Précisons que \hat{n} diminue strictement quand ψ_A augmente et

$$\hat{n} - \hat{T}_A + \frac{\sqrt{n}}{Ns} t_{1-\alpha/2, n_A-1} = \hat{n} \text{ à mesure que } \psi_A \text{ diminue;}$$

$$\hat{n} = \hat{T}_A + \frac{\sqrt{n}}{Ns} \left(\frac{1 + \hat{P}_A\hat{\sigma}_A^2/s^2}{2} \right)^{1/2} t_{1-\alpha/2, n_A-1} = \hat{n}_2 \text{ pour } \psi_A = 1.$$

et

$$\hat{n} - \hat{T}_A + \frac{\sqrt{n}}{Ns} \left(\frac{\sqrt{\hat{P}_A}\hat{\sigma}_A}{s} \right) t_{1-\alpha/2, n_A-1} = \hat{n}_3$$

quand ψ_A augmente.

On peut traiter la valeur du seuil inférieur d'une manière analogue dans chaque cas, ce qui donne trois intervalles de confiance concurrents, à savoir $IC_1(1 - \alpha) = (\hat{\ell}_1, \hat{n}_1)$, $i = 1, 2, 3$, où $\hat{\ell}_i$ a à peu près la même définition que \hat{n}_i dans (8), mais où $t_{1-\alpha/2, n_A-1}$ a été remplacé par $t_{\alpha/2, n_A-1}$. Les intervalles de confiance concurrents sont étiquetés par ordre de grandeur décroissant.

Le premier cas revient à supposer que σ_A^2 a une valeur élevée par rapport à $\text{var}(\hat{P}_A)$ et aboutit à la variance inconditionnelle usuelle, mais avec $n_A - 1$ degrés de liberté. Cela semble raisonnable dans la plupart des problèmes pratiques car σ_A^2 est une constante inconnue et $\text{var}(\hat{P}_A)$ correspond à $O(P_A/n)$. Le deuxième intervalle suppose qu'on adopte une distribution a priori normale tel qu'indiqué plus haut où $\sigma_A^2 = \hat{P}_A(1 - \hat{P}_A)/n$. Enfin, le dernier repose sur l'hypothèse que P_A dégénère essentiellement en \hat{P}_A .

2.5 Étude empirique pour l'EAS

Nous avons comparé plusieurs intervalles de confiance de la partie 2.4 dans le cadre d'une petite étude empirique,

Estimateurs du domaine

$$\hat{N}_A = \hat{p}_A^* N, \quad \hat{\mu}_A = \sum_{i=1}^{n_A} x_i' / n_A = \hat{T}_A / \hat{N}_A \text{ (défini seulement pour } n_A \geq 1), \text{ et } \hat{\sigma}_A^2 = \sum_{i=1}^{n_A} (x_i' - \hat{\mu}_A)^2 / (n_A - 1) \text{ (défini seulement pour } n_A \geq 2).$$

Il est subseqüemment entendu que $n_A \geq 2$ (ou, ce qui revient au même, que $\hat{p}_A \geq 2/n$), sauf indication contraire. À $n_A = 1$ ou 0, il vaut mieux coller l'étiquette «données insuffisantes» qu'essayer l'inférence. Les relations qui suivent dérivent directement des définitions.

$$T_A = N \hat{p}_A \hat{\mu}_A \text{ et } \hat{T}_A = N \hat{p}_A^* \hat{\mu}_A, \\ \bar{X} = \hat{p}_A \hat{\mu}_A \text{ et } \bar{x} = \hat{p}_A^* \hat{\mu}_A, \\ S_A^2 = \hat{p}_A (1 - \hat{p}_A) \hat{\mu}_A^2 + \hat{p}_A^* \hat{\sigma}_A^2$$

et

$$s_A^2 = \frac{n}{n-1} \hat{p}_A (1 - \hat{p}_A) \hat{\mu}_A^2 + \frac{n-1}{n \hat{p}_A - 1} \hat{\sigma}_A^2. \quad (1)$$

Il est assez facile de s'assurer que

$$\left(\sqrt{n/N} \right) \left(\hat{T}_A - T_A \right) = \sqrt{n \hat{\mu}_A} (\hat{p}_A - p_A) + \sqrt{\hat{p}_A} \sigma_A Z, \quad (2)$$

où $Z = \sqrt{n \hat{p}_A} (\hat{\mu}_A - \mu_A) / \sigma_A$. Sous réserve de \hat{p}_A , \hat{T}_A est donc biaisé pour T_A . Si, par exemple, on suppose la normalité sous-jacente et normalise $(\sqrt{n/N})(\hat{T}_A - T_A)$ par la variance conditionnelle correspondante, on obtient une distribution t non-centrale pour laquelle on ne connaît pas le paramètre de non-centralité proportionnel à $\sqrt{n \hat{\mu}_A} (\hat{p}_A - p_A)$, si bien qu'on ne dispose pas d'une base très solide pour effectuer des inférences (conditionnelles) valables. Tel est le problème que nous essayerons de résoudre dans les parties qui suivent.

Notons qu'en estimant la moyenne μ_A avec $\hat{\mu}_A$, le biais est égal à zéro et le problème précité ne se présente pas. C'est pourquoi on peut recourir à la méthode d'inférence type avec l'échantillonnage aléatoire simple, du moins lorsque les variables du domaine ont une distribution normale.

2.3 Méthode générale des intervalles de confiance

Soit $\hat{\theta} = (\hat{T}_A - T_A) / s_{T_A}^*$, où $s_{T_A}^2$ désigne un estimateur (à spécifier) de la variance (conditionnelle ou inconditionnelle) du total. Supposons qu'on connaisse la forme de la fonction de distribution conditionnelle (pour \hat{p}_A) de $\hat{\theta}$, par exemple $H(\cdot | \hat{p}_A; p_A, \mu_A, \sigma_A^2)$, où p_A, μ_A et σ_A^2 représentent des paramètres inconnus. Pour bâtir un intervalle de confiance (IC) conditionnel aux extrémités égales $(1 - \alpha) \times 100\%$ pour T_A , on établit un seuil maximal

$$c_n^* \equiv c_n^*(\alpha, \hat{p}_A, p_A) = -\inf\{x | H(x | \hat{p}_A; p_A) \geq \alpha/2\} = -H^{-1}(\alpha/2, \hat{p}_A; p_A)$$

où p_A est fixe et où on élimine temporairement la dépendance à μ_A et à σ_A^2 ; on définit seuil minimal (appelons-le c_A^0) de la même manière. IC $(1 - \alpha) = (\ell, u)$, donne un IC conditionnel aux extrémités égales $(1 - \alpha) \times 100\%$ pour T_A où

$$n = \hat{T}_A + c_n^* s_{T_A}^* \text{ et } \ell = \hat{T}_A + c_A^0 s_{T_A}^*. \quad (3)$$

Le problème pratique manifeste est que les seuils critiques c_n^* et c_A^0 ne dépendent pas seulement de \hat{p}_A mais aussi de p_A qu'on ne connaît pas. Une façon de contourner la difficulté consiste à adopter l'approche bayésienne et à supposer que p_A est la valeur réelle d'une variable aléatoire. En corrigeant la notation pour refléter l'hypothèse que p_A est stochastique, on remplace $H(x | \hat{p}_A; p_A)$ par $H(x | \hat{p}_A, p_A)$ et obtient

$$\Pr\{\theta \leq x | \hat{p}_A\} = F(X | \hat{p}_A)$$

$$= \frac{1}{\int h(\hat{p}_A') \int H(x | \hat{p}_A, p_A) f(\hat{p}_A | p_A) g(p_A) dp_A, \quad (4)$$

où $h(\hat{p}_A) = \int f(\hat{p}_A | p_A) g(p_A) dp_A$ et $g(p_A)$ correspond à la densité de p_A . À cause du plan d'échantillonnage, la distribution de $n \hat{p}_A$, conditionnelle à p_A , donne la fonction binomiale (n, p_A) , si bien que $f(\hat{p}_A | p_A)$ est connue. Avec l'approche bayésienne, les seuils c_n^* sont $c_n^* \equiv c_n^*(\alpha, \hat{p}_A) = -F^{-1}(\alpha/2 | \hat{p}_A)$ et $c_A^0 \equiv c_A^0(\alpha, \hat{p}_A) = -(1 - \alpha/2 | \hat{p}_A)$ de sorte que les limites supérieure et inférieure d'un IC conditionnel de $(1 - \alpha) \times 100\%$ pour T_A deviennent

$$n = \hat{T}_A + c_n^* s_{T_A}^* \text{ et } \ell = \hat{T}_A + c_A^0 s_{T_A}^*. \quad (5)$$

Aux fins qui nous intéressent, on suppose que la distribution a priori $g(p_A)$ est $N(\mu_{p_A}, \sigma_{p_A}^2)$ et qu'il faut spécifier μ_{p_A} et $\sigma_{p_A}^2$, sachant que $\sigma_{p_A}^2$ est assez faible pour que p_A se situe presque à coup sûr entre 0 et 1. On pose l'hypothèse de la normalité pour faciliter les calculs. Cette hypothèse saisit aussi la notion d'éventuels degrés de proximité et de symétrie avec μ_{p_A} . Pour la méthode empirique de Bayes, on utilise $\mu_{p_A} = \hat{p}_A$; plusieurs possibilités sont examinées pour $\sigma_{p_A}^2$ et analysées en détail ci-dessous. Nous avons appris par expérience que l'hypothèse de la normalité n'est pas essentielle; il ne s'agit que d'une question de commodité.

2.4 Intervalles de confiance avec des hypothèses normales

Pour continuer, supposons que les valeurs x_i du domaine A ont pour distribution $N(\mu_A, \sigma_A^2)$. Il se peut que l'hypothèse ne se vérifie pas dans la pratique. Quoiqu'il en soit, elle aboutit à des modifications qui ne donneront jamais de couverture inférieure à celle de l'approche habituelle pour les intervalles de confiance. Si on associe cette hypothèse aux résultats précédents, en particulier l'équation (2), et si on néglige les termes de degré inférieur, on obtient

L'article se présente comme suit. La partie 2 introduit les concepts. Nous examinons le cas des totalisations avec un échantillonnage aléatoire simple et nous en servons pour illustrer l'approche habituelle à l'estimation du domaine, le problème de couverture qui en résulte et l'approche adoptée pour y remédier. La partie 3 décrit l'extension de cette méthode à l'échantillonnage aléatoire stratifié et la partie 4 sert de conclusion.

2. ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

2.1 Méthode type

Särndal, Swensson et Wretman (1992; parties 3.3, 5.8, et chapitre 10) (par la suite SSW) exposent bien la méthode classique à l'estimation d'un domaine. La méthode qu'ils décrivent est générale. Nous la reproduisons ici pour un échantillonnage aléatoire simple et pour l'échantillonnage aléatoire stratifié, après une légère extension, en nous concentrant sur le total du domaine.

Soit x_i , la valeur de la caractéristique à laquelle on s'intéresse pour le i -ième ($i = 1, 2, \dots, N$) élément de la population, et soit A , un domaine auquel on s'intéresse. Nous envisagerons le cas où la base de sondage ne permet pas l'identification des éléments de A et où le nombre N_A d'éléments que renferme A est inconnu; SSW analysent à fond le cas où on connaît N_A . On suppose qu'il est possible d'identifier chaque élément de A inclus dans un échantillon. Le problème consiste à construire un intervalle de confiance pour le total du domaine $T_A = \sum_{i \in A} x_i$, d'après un échantillon de n éléments prélevés à même la base de sondage complète.

Que ce soit de façon implicite (comme à la partie 3.3 de SSW) ou explicite (comme à la partie 10.3), l'approche classique à la résolution du problème consiste à redéfinir x_i , en établissant $x_i = 0$ si $i \notin A$, ce qui contraint le total $T = \sum_{i=1}^N x_i$ à être égal au total T_A . On ne construit donc plus un intervalle de confiance pour le total d'un domaine mais pour un total de la population. Par la suite, nous supposons que les éléments x_i ont été redéfinis de la manière indiquée. Nous présumerons aussi dans l'article que n est assez important et n/N assez faible pour qu'on néglige les termes du deuxième degré. Définissons les autres paramètres de la population;

$$\bar{X} = T/N = \text{moyenne de la population};$$

$$S^2 = \sum_{i=1}^N (x_i - \bar{X})^2 / N = \text{variance de la population et } p_A = N_A / N = \text{proportion de la population se retrouvant dans } A.$$

Il découle que

$$(1) \quad \bar{p}_A = (N/n) \sum_{i=1}^n x_i / n, \quad \bar{x} = \sum_{i=1}^n x_i / n = \bar{p}_A / N, \quad s^2 = \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1), \quad \text{et } \bar{p}_A = n_A / n \text{ (où } n_A \text{ indique le nombre d'éléments de l'échantillon dans } A) \text{ ne sont pas biaisés pour les paramètres correspondants de la population;}$$

Paramètres du domaine

Définissons les paramètres et estimateurs que voici:

2.2 Définitions et notation

La proportion de la population qui se retrouve dans A est égale à $1 - p_A$ et $x_i = 0$ pour $i \in A$; par conséquent, quand p_A est petit et que les valeurs de x_i pour $i \in A$ sont très différentes de zéro, la distribution de x_i converge lentement. La distribution de $\sqrt{n}(\bar{T}_A - T_A)/N_s$ peut donc s'écarter de la normale, même pour des valeurs de n qu'on qualifierait habituellement de moyennes à élevées. On en trouvera une illustration avec la simulation de la partie 2.5. L'établissement de la couverture de l'intervalle de confiance pour les quantités du domaine au moyen des méthodes classiques peut aboutir à de piètres résultats quand l'échantillonnage aléatoire est stratifié. Dorfman et Valliant (1993) ont signalé le problème dans leur analyse de la distribution des salaires au sein de domaines composés de travailleurs appartenant à des professions précises. Le travail empirique préliminaire effectué par ces auteurs révèle que l'intervalle de confiance présumé de 95 %, établi pour le nombre total de travailleurs et la rémunération totale, pour les domaines de travailleurs mentionnés ci-dessus, n'offre en réalité qu'une couverture de 75 % à 85 %, même avec un échantillon global important ($n = 353$ établissements). Nous confirmons en partie ces résultats dans nos propres travaux empiriques, présentés à la partie 3. Par ailleurs, les auteurs précités ont constaté que la distribution de $\bar{T}_A - T_A$ dépend fortement de la valeur réelle de n_A , ce qui donne à penser qu'on devrait recourir à un intervalle de confiance conditionnel quelconque. Il semblerait souhaitable d'élaborer une méthode afin de construire des intervalles de confiance (conditionnels à n_A ou, ce qui est équivalent, à \bar{p}_A), pour T_A , de manière à assurer une couverture nominale ou presque, peu importe la taille réelle de l'échantillon du domaine. SSW abordent la question des inférences conditionnelles à la taille de l'échantillon à la partie 10.4, mais seulement quand on connaît N_A ; nous nous intéresserons ici au cas où cette valeur est inconnue.

Il s'ensuit que $\sqrt{n}(\bar{T}_A - T_A)/(N_s) \xrightarrow{d} N(0, 1)$. Donc, quand n est assez important, on peut se servir des valeurs appropriées de la distribution normale pour construire les intervalles de confiance de T_A , comme l'indiquent SSW, p. 391.

- (2) $E(\bar{T}_A) = T_A,$
- (3) $\text{var}(\bar{T}_A) = N^2 S^2 / n,$
- (4) $\sqrt{n}(\bar{T}_A - T_A)/(N_s) \xrightarrow{d} N(0, 1),$ et
- (5) s^2 est convergent pour $S^2.$

Intervalle de confiance des paramètres de domaine quand la taille de l'échantillon du domaine est aléatoire

ROBERT J. CASADY, ALAN H. DORFMAN et SUOJIN WANG¹

RÉSUMÉ

Soit A , le domaine de la population auquel on s'intéresse. Supposons qu'il est impossible d'identifier les éléments de A dans la base de sondage et qu'on ignore le nombre d'éléments que contient A . Supposons en outre qu'on prélève un échantillon de taille fixe (n par exemple) de la base de sondage et que la taille de l'échantillon du domaine résultant (appelons-la n_A) soit aléatoire. Le problème consiste à bâtir un intervalle de confiance pour un paramètre du domaine tel que l'agrégat du domaine $T = \sum_{i \in A} x_i$. Habituellement, la solution consiste à redéfinir x_i en établissant $x_i = 0$ si $i \notin A$. Au lieu de construire un intervalle de confiance pour le total du domaine, on en construit donc un pour un total de la population, ce que permet de satisfaire la théorie de la distribution normale (de façon asymptotique pour n). Une autre solution consisterait à imposer des conditions à n_A et à bâtir des intervalles de confiance à couverture presque nominale, avec certaines hypothèses se rapportant à la population du domaine. Les auteurs évaluent la nouvelle approche de manière empirique au moyen de populations artificielles et des données de l'Occupational Compensation Survey du Bureau of Labor Statistics (BLS).

MOTS CLÉS: Méthodes de Bayes; conditionnement; enquêtes auprès des établissements; échantillonnage aléatoire simple; stratification; méthodes d'enquête.

1. INTRODUCTION

valable, l'écart-type est considéré comme une source d'erreur.

Lorsqu'on estime un domaine, les intervalles de confiance construits de la manière classique peuvent donner lieu à une couverture nettement insuffisante, aspect que néglige parfois la littérature. C'est ce que nous appelons le «problème du domaine». Nous l'examinerons en faisant appel à une méthode assez complexe articulée sur les principes bayésiens. Cette méthode aboutit cependant à une solution pratique assez simple, qui améliore la méthode-logie présentement en usage. La principale distinction est que la nouvelle méthode utilise une statistique t de Student dont le nombre de degrés de liberté dépend du nombre et de la configuration des éléments que renferme le domaine dans l'échantillon, au lieu de la statistique normale type, pour construire les intervalles de confiance.

Nous nous concentrerons sur les totaux et les moyennes du domaine pour les deux cas courants que sont l'échantillonnage aléatoire simple et l'échantillonnage aléatoire stratifié. Dans le premier cas, les méthodes habituelles donnent des résultats satisfaisants pour la moyenne; la couverture du total peut néanmoins être inférieure à la couverture nominale, mais généralement pas d'une manière préoccupante. Avec l'échantillonnage aléatoire stratifié, les intervalles de confiance soulèvent de sérieuses difficultés à l'égard du niveau de couverture, tant pour la moyenne que pour le total. Dans ce cas, on ajoute à la nouvelle méthodologie une approximation très connue développée par Satterthwaite (1946). Des approches différentes de la nôtre, mais utilisant la même approximation peuvent être étudiées dans Johnson et Rust (1993) et Kott (1994).

Ceux qui échantillonnent une population finie désirent souvent estimer des totaux, des moyennes ou d'autres quantités de certaines parties de la population en question appelées habituellement «domaines». La base de sondage n'énumère pas explicitement ces domaines, on ne connaît pas d'avance le nombre d'items que comprendra l'enquête et le nombre d'éléments qu'on retrouvera dans la population est plus que souvent lui aussi inconnu. Ainsi, on pourrait vouloir échantillonner des élèves à l'égard de certains problèmes de santé, puis connaître la pression sanguine moyenne des enfants du groupe dont le poids est inférieur à la normale. Ces derniers constitueraient un domaine. Or, le seul renseignement dont on dispose pour déterminer si un enfant est sous-alimenté ou non se trouve vraisemblablement parmi les enfants échantillonnés; dans un tel cas, le domaine ne fait pas explicitement partie de la base de sondage. Estimer la précision des estimateurs forme une part capitale du processus d'inférence. On y parvient typiquement en estimant l'écart-type, le coefficient de variation ou l'intervalle de confiance. Peu importe la mesure utilisée pour établir la précision, elle suppose un intervalle de confiance valable. De par leur construction, tous les intervalles de confiance comprennent un niveau de confiance «nominal». Pour être valable, l'intervalle de confiance doit avoir une couverture réelle identique à la couverture nominale. On peut calculer la couverture réelle de façon théorique ou empirique, en reproduisant les circonstances dans lesquelles l'intervalle de confiance serait utilisé dans la pratique. S'il ne donne pas un intervalle de confiance

¹ Robert J. Casady et Alan H. Dorfman, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave. N.E., Washington D.C., 20212-0001, U.S.A.; Suojin Wang, Department of Statistics, Texas A&M University, College Station, TX 77843, U.S.A.

- GODAMBE, V.P., et THOMPSON, M.E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Revue Internationale de Statistique*, 54, 127-138.
- LEHTONEN, R., et PAHKINEN, E.J. (1996). *Practical Methods for Design and Analysis of Complex Surveys*. Edition révisée. Chichester: John Wiley & Sons.
- LEHTONEN, R., et VEIJANEN, A. (1998). On Multinomial Logistic Generalized Regression Estimators. Jyväskylä: Preprints from the Department of Statistics, University of Jyväskylä, 22.
- McCULLAGH, P., et NELDER, J.A. (1989). *Generalized Linear Models*. Deuxième édition. London: Chapman and Hall.
- NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics*, 5, 223-239.
- SKINNER, C.J., HOLT, D., et SMITH, T.M.F. (Eds) (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Pour la stratification a posteriori incomplète, ou la procédure itérative, le modèle d'analyse de la variance des effets principaux reposait sur des variables auxiliaires classées. Nous avons comparé les modèles avec et sans l'indicateur de recherche d'un emploi. Le troisième modèle comprenait un polynôme du quatrième degré pour l'âge.

4.3.2 Résultats

Sans les variables auxiliaires, les estimateurs HT aboutissent habituellement à une variance plus élevée que les estimateurs de régression généralisés (tableau 2). Les résultats sont légèrement meilleurs pour ces derniers que pour les estimateurs HT quand on recourt à une procédure itérative combinant l'âge, le sexe et la région. Les résultats étaient nettement supérieurs avec les modèles intégrant l'indicateur de recherche d'un emploi, qui présente une corrélation plus étroite ($r = 0,83$) avec l'indicateur du chômage du BIT que les autres variables auxiliaires. Les variables auxiliaires améliorent donc la précision de l'estimation (voir Djert 1997).

Tableau 2

Propriétés des estimations du taux de chômage (T (%)) selon la procédure itérative du quotient (R) et le modèle incluant le polynôme pour l'âge (P), avec (B) ou sans (N) indicateur de recherche d'un emploi. $E-T$, désigne l'écart-type et (TC) (%), le taux de couverture à l'intervalle de confiance de 95 %

Modèle	Méthode	T	Biais	$E-T$	Deff	TC	EMARD
HT		20,32	-0,0081	1,461	1	95,7	35,28
RN	REGG	20,30	-0,0262	1,454	0,995	95,3	46,03
RN	REGGL	20,31	-0,0229	1,454	0,995	95,3	45,93
RE	REGG	20,30	-0,0244	0,895	0,612	96,0	35,74
RE	REGGL	20,29	-0,0419	0,901	0,617	94,8	34,80
PE	REGG	20,30	-0,0259	0,887	0,607	95,6	35,41
PE	REGGL	20,29	-0,0421	0,896	0,613	95,1	34,76

Tableau 3

Erreur moyenne absolue relative au domaine (EMARD) et taux de couverture moyen (TC) (%) des intervalles de confiance à 95 % pour la fréquence estimée des classes des domaines dont la fréquence réelle $f_{(d)}^i$ ($i = 1, 2, 3$) a) est inférieure à 100, et b) est au moins égale à 100. Le modèle inclut le polynôme pour l'âge

Méthode	$\hat{f}_{(d)1}$			$\hat{f}_{(d)2}$			$\hat{f}_{(d)3}$			TC
	REGG	REGGL	EMARD	REGG	REGGL	EMARD	REGG	REGGL	EMARD	
a)	96,92	80,28	67,20	121,95	88,2	77,8	84,6	93,7	93,9	93,3
b)	6,95	12,31	14,35	94,1	85,9	93,7	93,7	85,4	93,9	93,3

Les estimations REGG et REGGL présentent peu de variations au niveau de la population (tableau 2). La méthode REGGL ne s'est jamais avérée inférieure à la

méthode REGG. Elle a donné des estimations plus précises des totaux des domaines que la seconde (tableau 3). Quand le modèle inclut l'âge sous forme de variable auxiliaire continue, l'écart-type du taux de chômage estimé est plus faible avec la méthode REGGL qu'avec la méthode REGG dans 19 domaines sur 20. Malheureusement, les intervalles de confiance calculés avec REGGL sont souvent trop étroits à cause des faibles estimations de la variance (tableau 3).

5. RÉCAPITULATION

Nous proposons une nouvelle approche à l'estimation de la fréquence des classes au sein d'une population au moyen d'un modèle, pour une variable discrète associée aux réponses dans le cadre d'une enquête par sondage. Notre méthode d'estimation généralisée par régression logistique (REGGL) repose sur un modèle logistique multinomial qui pourrait s'avérer plus réaliste à l'égard des indicateurs de classe que le modèle linéaire dont on se sert normalement pour l'estimation de régression généralisée (REGG). Les estimateurs REGGL et REGG aboutissent à des résultats identiques avec la stratification a posteriori complète, mais les résultats diffèrent avec d'autres modèles comme la procédure itérative. Comparativement à REGG, la méthode REGGL requiert habituellement plus d'information auxiliaire, et pas seulement les totaux auxiliaires. Quoiqu'il en soit, elle semble préférable à la méthode REGG quand la probabilité des classes varie considérablement avec la fourchette des variables auxiliaires continues, et lorsqu'on a besoin d'estimations pour de petits domaines, en particulier si la fréquence des classes est peu élevée.

REMERCIEMENTS

Les auteurs remercient le professeur Carl-Erik Särndal, de l'Université de Montréal, pour ses commentaires sur la version antérieure de l'article. Les remarques détaillées d'un rédacteur associé et de deux examinateurs nous ont été d'une grande utilité. Nous nous devons aussi de remercier M. Timo Koskimäki, de Statistics Finland, pour nous avoir fourni les données de l'Enquête sur la population active, ainsi que M. Kari Djert, pour ses commentaires judicieux.

BIBLIOGRAPHIE

CHAMBERS, R.L. (1996). Robust case-weighting for multipurpose establishment surveys. *Journal of Official Statistics*, 12, 3-32.

DJERT, K. (1997). Effects of post-stratification on the estimates of the Finnish Labour Force Survey. *Journal of Official Statistics*, 13, 29-39.

ESTEVAO, V., HIDROGLOU, M.A., et SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.

4.3 Application aux données de l'Enquête sur la population active finlandaise

4.3.1 Population artificielle

Nous avons examiné l'estimation du taux de chômage au

moyen des données de l'Enquête sur la population active (EPA) finlandaise couvrant trois mois successifs de 1994. Cette population comprenait 33 329 sujets. Le Registre de la population nous a permis d'établir le groupe d'âge (15-24, 25-34, 35-44, 45-54 et 55-64 ans), le sexe et la région (trois régions) de chaque sujet. Nous avons aussi tiré un indicateur recherche d'un emploi du registre que garde le ministère du Travail indiquant, quelles personnes au chômage recherchent un emploi. Cette source de données administratives se caractérise par un décalage d'environ deux semaines. La proportion de personnes dont la situation réelle sur le marché du travail change au cours d'un si bref laps de temps devrait être relativement faible. Il convient de souligner que le statut de personne à la recherche d'un emploi dans le registre n'a pas la même définition que dans l'Enquête sur la population active. Dans cette dernière, la mesure repose sur la définition normalisée par le Bureau international du travail (BIT). Toutes les données auxiliaires ont été amalgamées aux données d'enquête, pour chaque sujet.

Le taux de non-réponse varie avec le statut recherche d'un emploi. Ainsi, il s'établissait à 1,4 % pour les personnes inscrites à la recherche d'un emploi et à 7,6 % pour les autres. On a modélisé la probabilité de non-réponse au moyen d'un modèle d'analyse de la variance logistique et les estimations du taux de non-réponse le plus vraisemblable (variant de 2,9 % à 22,8 %) ont servi de modèle de non-réponse dans les simulations.

Pour les simulations, nous avons bâti une population artificielle de $N = 30\ 835$ personnes. La situation d'emploi pouvait être de trois sortes: "occupé", "au chômage" et "pas dans la population active" dont la fréquence, au sein de la population, s'établissait respectivement à $t_1 = 17\ 373$, $t_2 = 4\ 433$, et $t_3 = 9\ 029$. Le taux de chômage était défini par $R = t_2 / (t_1 + t_2) = 20,33\ %$. Comme domaine, nous avons utilisé les cellules des totalisations croisées selon le groupe d'âge, le sexe et la situation d'emploi notée au registre.

De la population artificielle, on a prélevé $K = 1\ 000$ échantillons indépendants de $n = 1\ 000$ (EASSR). Le modèle de non-réponse a été ajusté à la population originale pour simuler la non-réponse dans chaque échantillon. Ensuite, on a estimé la probabilité de réponse de chaque échantillon par régression logistique en recourant au même modèle d'analyse de la variance que pour le modèle de non-réponse. Enfin, nous avons multiplié chaque probabilité π_k par la probabilité de réponse estimée. La comparaison de REGGL et REGG repose sur trois modèles. Les composantes de x_k étaient des variables nominales correspondant à l'âge (5 groupes), au sexe, à la région (3 régions) et au statut "recherche d'un emploi".

4.2 Essai avec les données simulées

Dans les estimations REGG (2), la variance correspondait à une constante $\sigma_{\pi}^2 = \sigma^2$, dont les valeurs se sont annulées. Avec REGGL, on a estimé la fréquence des domaines au moyen de l'équation (5) et la variance avec l'équation (7). Pour la méthode REGG et l'estimateur HT, lire Särndal et coll. (1992, p. 401). On a calculé les intervalles de confiance des fréquences comme si les indicateurs de classe étaient indépendants. Au seuil de signification nominal de 95 %, le taux de couverture acceptable se situe à l'intérieur de [93,65 %, 96,35 %] pour $K = 1\ 000$ échantillons simulés.

$$\text{EMARD}(i) = \frac{1}{D} \sum_{p=1}^D \frac{1}{K} \sum_{j=1}^K \frac{100 \left| \hat{t}_{(d_p)}^{(s_j)} - t_{(d_p)}^{(s_j)} \right|}{t_{(d_p)}^{(s_j)}}.$$

Monte Carlo de l'estimateur HT (Lehtonen et Pakkinen 1996). Nous avons déterminé la précision globale des estimations relatives aux domaines grâce à l'erreur moyenne absolue relative au domaine (EMARD), pour D domaines et K échantillons s_j :

Pour comparer REGGL et REGG, nous avons simulé un ensemble de données dans lequel la variable auxiliaire X correspondait à une variable aléatoire continue, distribuée uniformément dans $(-3,3)$. La variable Y , à laquelle on s'intéresse, représentait trois classes suivant la distribution spécifiée par $x_k' \beta_1 = 0$, $x_k' \beta_2 = 3X_k - 1$ et $x_k' \beta_3 = -2X_k$ pour $N = 10\ 000$ éléments ($k = 1, 2, \dots, N$). On a prélevé un millier d'échantillons de taille $n = 1\ 000$ indépendamment par EASSR. X_k et X_k^2 servaient de variables auxiliaires. Aucun estimateur ne semble biaisé (tableau 1). Par ailleurs les estimations de la variance présentaient un biais empirique inférieur à 3 % et un écart-type de moins de 5 %.

Tableau 1

Effets de plan d'échantillonnage (Deff) sur les estimateurs de la fréquence des classes et taux de couverture empiriques (TC) (%) à l'intervalle de confiance de 95 % pour les classes $i = 1, 2, 3$

Méthode	Deff			TC		
	\hat{t}_1	\hat{t}_2	\hat{t}_3	\hat{t}_1	\hat{t}_2	\hat{t}_3
HT	1	1	1	95,2	95,3	94,7
REGG	0,93	0,55	0,57	95,0	94,3	95,6
REGGL	0,89	0,45	0,50	94,9	93,7	95,3

Les meilleurs résultats sont ceux obtenus avec la méthode REGGL, vraisemblablement parce que les fréquences proportionnelles des classes varient considérablement pour la gamme des variables auxiliaires. La probabilité de chaque classe dépendait tellement de la variable auxiliaire continue que le modèle de régression linéaire s'ajustait mal aux données.

$$\hat{t}_i = \sum_{k \in U} \hat{p}_{ki} + \sum_{k \in S} a^k(z_{ki} - \hat{p}_{ki}) \quad (i = 1, 2, \dots, m). \quad (4)$$

Les estimateurs REGG et REGGL en (3) et (4) incluent une somme des valeurs prédites pour la population. En

réalité cependant, on n'a pas vraiment besoin de connaître x_k pour chaque élément de la population U . Avec REGG (3), le total des variables auxiliaires $\sum_{k \in U} x_k$ suffit, car (3) peut aussi prendre la forme $\hat{t}_G = \hat{t}_{HT} + (\sum_{k \in U} x_k - \sum_{k \in S} a^k x_k) \hat{b}_i$.

Dans le cas particulier d'une stratification a posteriori complète, l'information requise pour REGGL est similaire à celle nécessaire pour REGG. Dans les autres cas, notamment celui de la stratification a posteriori incomplète, on ne peut calculer $\sum_{k \in U} \hat{p}_{ki}$ en (4) sans connaître la fréquence de chaque valeur x_k dans la population. Avec deux variables auxiliaires, par exemple, on a besoin de la fréquence des valeurs marginales pour REGG, mais pour REGGL, il faut la fréquence des cellules.

On ne procède pas à des estimations que pour l'ensemble de la population, on estime aussi les sous-populations. La population U est divisée en domaines $U^{(p)} \subset U$ de taille $N^{(p)}$. L'ensemble s de répondants se compose des sous-ensembles $s^{(p)} = s \cap U^{(p)}$ comprenant $n^{(p)}$ éléments. Comme pour l'estimation REGG (Särndal et coll. 1992), on applique l'estimateur REGGL

$$\hat{t}^{(d)i} = \sum_{k \in U^{(d)}} \hat{p}_{ki} + \sum_{k \in S^{(d)}} a^k(z_{ki} - \hat{p}_{ki}). \quad (5)$$

Les estimateurs s'additionnent: $\sum_i \hat{t}^{(d)i} = N^{(d)}$. Quand on réunit deux domaines d_1 et d_2 , qui ne se chevauchent pas, l'estimation REGGL de $d = d_1 \cup d_2$ correspond à $\hat{t}^{(d)i} = \hat{t}^{(d_1)i} + \hat{t}^{(d_2)i}$. Par conséquent, $\sum_d \hat{t}^{(d)i} = \hat{t}_i$ pour les domaines qui ne se chevauchent pas et $\sum_i \hat{t}_i = N$.

Dans l'estimation par régression généralisée, les résultats obtenus grâce à (3) ou (4) peuvent être négatifs quand les valeurs résiduelles négatives coïncident avec une valeur élevée de a^k . Les estimations REGG négatives se multiplient lorsque le nombre de variables auxiliaires augmente (Chambers 1996), ce qui ne se produit pas avec l'estimation REGGL, \hat{p}_{ki} étant borné par la formulation du modèle. Lors de nos essais, nous n'avons obtenu d'estimations REGGL négatives que dans quelques cas, et pour de petits domaines. Les estimations REGGL correspondent souvent à la somme des probabilités estimées, si bien qu'elles sont toujours positives (voir la partie 3.2).

Si le modèle (1) inclut une variable explicative auxiliaire, la méthode REGGL en estime exactement le total pour la population. Une propriété d'étalement de ce genre présente un grand intérêt pour maintes applications.

3.2 Estimation du maximum de vraisemblance au moyen du logarithme du rapport de vraisemblance pondéré par π

On estime le paramètre β du modèle (1) en maximisant le logarithme du rapport de vraisemblance pondéré par π

$$L_s(\beta^1, \dots, \beta^m) = \sum_{k \in S} \pi_k^{-1} \left\{ I\{Y^k = 1\} \log \left(1 - \sum_{m=2}^l \mu_{ki} \right) + \sum_{m=2}^l I\{Y^k = i\} \log \mu_{ki} \right\}$$

(Godambe et Thompson 1986; Nordberg 1989; Särndal et coll. 1992, p. 517). En général, on maximise la fonction de probabilité numériquement au moyen des méthodes appropriées comme l'algorithme de Newton-Raphson. Avec une stratification a posteriori complète, on peut montrer que les valeurs rajustées de \hat{z}_{ki} de REGG correspondent aux estimations de \hat{p}_{ki} de REGGL. Par conséquent, les estimateurs REGG et REGGL sont identiques quand aucune cellule ne manque dans la stratification a posteriori (Lehtonen et Veijanen 1998). Cette remarque ne s'applique pas aux autres modèles comme celui de la stratification a posteriori incomplète.

L'estimateur REGGL (4) comporte deux éléments: une somme des probabilités estimées pour la population et un terme d'ajustement $\sum_{k \in S} a^k(z_{ki} - \hat{p}_{ki})$. Il est possible de montrer que si le modèle intègre une coordonnée à l'origine, ce terme disparaît et la fréquence t_i est estimée par $\sum_{k \in U} \hat{p}_{ki}$ (Lehtonen et Veijanen 1998). Nous nous sommes servis d'un estimateur par ratio $\hat{R} = \hat{t}_i / (\hat{t}_1 + \hat{t}_j)$, dont on estime la variance par les techniques de linéarisation de Taylor (Särndal et coll. 1992, p. 179):

$$V(\hat{R}) = \frac{1}{(\hat{t}_1 + \hat{t}_j)^2} \left[(1 - \hat{R})^2 \hat{C}_{ii} + 2\hat{R}(\hat{R} - 1) \hat{C}_{ij} + \hat{R}^2 \hat{C}_{jj} \right] \quad (6)$$

où C_{ij} , la covariance de \hat{t}_i et de \hat{t}_j est estimée par

$$\hat{C}_{ij} = \sum_{k \in S} \frac{\Delta_{kp} e_{ki} e_{kj}}{e_{pi} e_{pj}} \pi_k \pi_p. \quad (7)$$

Dans (7), $e_{ki} = z_{ki} - \hat{p}_{ki}$ et $\Delta_{kp} = \text{Cov}(I^k I^p) = \pi_{kp} - \pi_k \pi_p$. Des dérivations analogues sont valables pour les estimateurs des domaines correspondants.

4. ESSAIS

4.1 Description des simulations

Dans toutes les simulations, on a prélevé $K = 1\,000$ échantillons d'une population par échantillonnage aléatoire simple sans remise (EASSR). La moyenne et l'erreur-type de Monte Carlo des estimations ont été établies à partir des échantillons utilisés pour la simulation. L'effet du plan d'échantillonnage sur l'estimateur $\hat{t}^{(d)i}$ a été calculé comme un ratio des variances estimées: $\text{Deff}(\hat{t}^{(d)i}) = \hat{V}_{mc}(\hat{t}^{(d)i}) / \hat{V}_{HT}(\hat{t}^{(d)i})$, où $\hat{V}_{mc}(\hat{t}^{(d)i})$ représente la variance estimée de

linéaire, pour les petits domaines. Nous avons aussi obtenu de bons résultats avec une seule variable auxiliaire continue.

L'article se structure comme suit. La partie 2 expose le

modèle logistique multinomial et les principes fondamentaux. À la partie 3, on présente les estimateurs de régression généralisés pour la fréquence des classes dans une population et divers domaines, et parle de l'estimation des paramètres du modèle par le logarithme pondéré du rapport de vraisemblance. Suit la présentation des estimateurs de la variance. Les tests de Monte Carlo sont analysés à la partie 4, et la partie 5 sert de conclusion.

2. MODÈLE

Soit les variables discrètes aléatoires Y_k à m dimensions

associées à N éléments k dans une population finie U . Les valeurs réelles y_k ne sont observées que pour un échantillon $s \subset U$ de taille n . L'objectif consiste à estimer la distribution de fréquence des valeurs y_k au sein de la population, dans les problèmes de classification, on estime la proportion des classes. Supposons que le vecteur des variables auxiliaires x_k soit connu pour chaque membre de la population. Appliquons le modèle logistique multinomial

$$P\{Y_k=i\} = \frac{\exp\{x_k' \beta_i\}}{\sum_{j=1}^m \exp\{x_k' \beta_j\}} \quad (i = 1, 2, \dots, m) \quad (1)$$

et supposons que Y_k est conditionnellement indépendant, étant donné x_k . C'est ce modèle dont on se sert pour la régression logistique avec le cas binaire. Le vecteur des paramètres β se compose des vecteurs β_i ($i = 1, 2, \dots, m$) ayant pour composantes β_{ij} ($j = 1, 2, \dots, q$). On suppose que les paramètres peuvent être discernés, c'est-à-dire que deux valeurs paramétriques n'aboutiront pas à deux probabilités identiques (1) pour l'élément k . Il s'ensuit que les variables auxiliaires x_{kj} ($j = 1, 2, \dots, q$) sont linéairement indépendantes. Pour éviter des difficultés au niveau de l'identification, on établit que $\beta_1 = 0$. Il est facile de généraliser (1) de sorte qu'on peut attribuer des variables auxiliaires différentes aux m classes (Lehtonen et Veijanen 1998).

Le plan d'échantillonnage spécifie les probabilités d'inclusion des éléments de la population. La probabilité d'inclusion du k -ième élément prélevé est égale à π_k et les éléments k et p se retrouvent simultanément dans l'échantillon s avec la probabilité $\pi_{kp} > 0$ ($\pi_{kk} = \pi_k$). Comme d'habitude, on presume que les indicateurs d'appartenance à l'échantillon $I_k = I\{k \in s\}$ sont conditionnellement indépendants de Y_k , étant donné x_k , mais la probabilité d'inclusion pourrait illustrer une corrélation avec les variables auxiliaires.

Avec la non-réponse unitaire, si la probabilité que l'élément k réponde correspond à θ_k , indépendamment de I_p

et de X_p ($p \in U$), on peut remplacer $\pi_k \theta_k$ par π_k . Parallèlement, on remplace $\pi_{kp} \theta_k \theta_p$ quand les éléments répondent de façon indépendante les uns avec les autres.

3. ESTIMATION PAR RÉGRESSION GÉNÉRALISÉE LOGISTIQUE

3.1 Définition de REGGL

Pour estimer la distribution de fréquence de y_k , on définit les indicateurs de classe $Z_{ki} = I\{Y_k = i\}$ avec les réalisations z_{ki} et on estime les totaux $t_i = \sum_{k \in U} z_{ki}$. L'estimateur de Horvitz-Thompson (HT) de t_i correspond à $t_i^{HT} = \sum_{k \in s} a_k z_{ki}$, les poids de l'échantillon étant donnés par $a_k = 1/\pi_k$. Pour rendre l'estimation par régression généralisée (REGGL) plus facile, on recourt au modèle de régression $Z_{ki} = x_k' \beta_i + \varepsilon_{ki}$ avec $\text{Var}(\varepsilon_{ki}) = \sigma_{ki}^2$ (Särndal et coll. 1992; Estevao et coll. 1995). On estime le paramètre β_i par l'équation

$$\beta_i = \left(\sum_{k \in s} a_k \frac{x_k x_k'}{z_{ki}} \right)^{-1} \left(\sum_{k \in s} a_k \frac{x_k z_{ki}}{z_{ki}} \right) \quad (i = 1, 2, \dots, m) \quad (2)$$

et les valeurs rajustées $\hat{z}_{ki} = x_k' \beta_i$ sont intégrées à l'estimateur REGGL

$$t_G^i = \sum_{k \in U} \hat{z}_{ki} + \sum_{k \in s} a_k (z_{ki} - \hat{z}_{ki}) \quad (i = 1, 2, \dots, m). \quad (3)$$

Choisir un modèle linéaire pour l'estimateur REGGL (3) est partiellement justifié quand la variable de la réponse est continue. Pour les mesures binaires Z_{ki} toutefois, l'emploi d'un modèle linéaire peut sembler irréaliste. En général, on préférerait un modèle logistique. En effet, avec une formulation logistique, la valeur prévue se retrouve toujours à l'intérieur de $[0, 1]$, tandis qu'avec la formulation linéaire, elle peut dépasser ces limites naturelles. Si la probabilité que $Z_{ki} = 1$ approche 0 ou 1, les deux modèles donneront des résultats différents. En outre, quand on compte $m > 2$ classes, le bon sens veut qu'on décrive la distribution combinée de Z_{ki} ($i = 1, 2, \dots, m$) au moyen du modèle logistique multinomial (1). Pour appliquer ce modèle à l'estimation par régression généralisée, on estime les probabilités $\mu_{ki} = E(Z_{ki} | x_k; \beta) = P\{Y_k = i | x_k; \beta\}$ grâce à

$$p_{ki} = P\{Y_k = i | x_k; \beta\} = \frac{\exp\{x_k' \beta_i\}}{1 + \sum_{j=2}^m \exp\{x_k' \beta_j\}},$$

qui ne présente aucun lien linéaire avec les variables auxiliaires. L'estimateur de régression généralisé logistique (REGGL) se définit ainsi:

Estimeurs de rgression gnralses logistiques

RISTO LEHTONEN et ARI VEIJANEN¹

RSUM

Dans cet article, les auteurs examinent comment estimer la frquence des classes d'une variable discrte associe aux rponses par le biais d'un modle. L'estimation fait appel une nouvelle mthode d'estimation des donnes d'enqte, etroitement liee l'estimation par rgression gnralise. Dans cette dmarche, les donnes sur les variables auxiliaires sont intgrees la mthode d'estimation par ajustement au moyen d'un modle linera. Au lieu de recourir un modle linera pour les indicateurs de classe, nous dcrivons la distribution combinee des indicateurs de classe par un modle logistique multinomial. Les auteurs prsentent des estimeurs de rgression gnralses logistiques pour la frquence des classes au sein d'une population et de divers domaines. Ils ont entrepris des essais de Monte Carlo sur des donnes simules et les donnes relles issues de l'Enqte sur la population active menes chaque mois par Statistique Finland. L'estimation de rgression gnralisee logistique donne de meilleurs rsultats que l'estimation de rgression ordinaire pour les petits domaines, en particulier les classes faible frquence.

MOTS CLS: Information auxiliaire; frquence des classes; modles linera; gnralses; enqte sur la population active; estimation assiste par modle; estimeurs de rgression.

1. INTRODUCTION

Envisageons l'estimation de la frquence des classes d'une variable discrte associe une rponse dans le cadre d'une enqte par sondage. Le nombre de sujets dans la classe correspond la somme des indicateurs de la classe pour la population, bref au total de cet indicateur. On peut donc recourir des mthodes conues pour estimer la population totale afin de rsoudre le problme. Pour accroitre la prcision de l'estimation, le statisticien d'enqte fera souvent appel aux donnes auxiliaires existantes. Si la valeur probable de la variable associe la rponse prsente un lien linera avec les variables auxiliaires, comme cela peut se produire avec les variables continues, il vaut la peine de recourir l'estimateur de rgression gnralise (Särndal, Swensson et Wretman 1992; Estevao, Hidiroglou et Särndal 1995). En effet, la rgression gnralisee peut donner une meilleure estimation et attnuer le biais attribuable la non-rponse unitaire si les variables secondaires prsentent de fortes corrlations avec la variable principale de la rponse.

Du point de vue du concepteur, un modle linera s'avre fort restrictif et pourrait ne pas constituer le choix idal pour les variables binaires comme la situation d'emploi d'une personne (occupe, au chmage) ni pour les variables discrtes comme sa situation sur le march du travail (occupe, au chmage, inactive), d'une manire plus gnrale. Nous proposons pour ces variables un genre d'estimateur de rgression gnralise logistique s'inspirant du modle logistique multinomial qui dcrit la distribution combinee des indicateurs de classe. La raison pour laquelle ce modle a t retenu est donc identique celle voquee pour les modles linera gnralses (McCullagh et Nelder 1989).

Nous estimerons les paramtres du modle logistique en maximisant le logarithme du rapport de vraisemblance pondr d'un chantillon, soit l'estimateur de Horvitz-Thompson pour la fonction du rapport de vraisemblance de la population (Godambe et Thompson 1986; Nordberg 1989; Skinner, Holt et Smith 1989; Särndal et coll. 1992, p. 517).

Comme application, nous prendrons l'estimation du taux de chmage dans le cadre de l'Enqte sur la population active effectuee chaque mois par Statistique Finland. Les dossiers administratifs indiquant si une personne la recherche d'un emploi est inscrite au bureau d'emploi local peuvent servir de donnes auxiliaires d'un registre et ont t combines aux donnes d'enqte sur chaque sujet grce au numro d'identit personnel, unique dans chaque source de donnes. La variable auxiliaire correspondante prsente une troite corrlation avec les rsultats de l'enqte sur le chmage. L'usage de ces donnes administratives dans l'estimation devrait donc amliorer cette dmarche et attnuer le biais. D'autres donnes auxiliaires (sexe, ge, donnes rgionales) viennent du Registre de la population. Elles aussi ont t combines aux donnes d'enqte au niveau individuel.

Nous avons tudi les proprits des estimeurs de rgression gnralses avec les techniques de simulation de Monte Carlo en vertu desquelles des chantillons BASSR ont t prlevs de faon rptitives d'une population constitue partir des donnes de l'Enqte sur la population active. Nous avons recouru une stratification a posteriori incomplte ou une procdure itrative reposant sur un modle d'analyse de la variance des effets principaux. Les rsultats des essais rvlent que la formulation logistique donne de meilleurs rsultats que la formulation

¹ Risto Lehtonen et Ari Veijanen, Statistics Finland, P.O. Box 5A, FIN-00022 Statistics Finland, Finlande.

ANNEXE

Voici la preuve que la valeur optimale de Ω_{ij} est celle donnée en (3.4). La fonction de Lagrange s'écrit comme suit

$$L = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (\Omega_{ij} - D_{ij})^2 -$$

$$2\lambda \left[\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \Omega_{ij} (d'_i x_i - d_j x_j)^2 - V^{YG} (X_{HT}) \right]. \quad (A.1)$$

En dérivant (A.1) par rapport à Ω_{ij} et en rendant l'expression égale à zéro, on obtient

$$\Omega_{ij} = D_{ij} + \lambda D_{ij} \bar{\sigma}_{ij} (d'_i x_i - d_j x_j)^2. \quad (A.2)$$

En insérant (A.2) dans (3.2), l'expression devient

$$\lambda = \frac{V^{YG} (X_{HT}) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d'_i x_i - d_j x_j)^2}{\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} \bar{\sigma}_{ij} (d'_i x_i - d_j x_j)^4}. \quad (A.3)$$

Il suffit de remplacer (A.3) dans (A.2) pour obtenir la valeur optimale de Ω_{ij} qui apparaît en (3.4).

BIBLIOGRAPHIE

- BRATLEY, P., FOX, B.L., et SCHRAGE, L.E. (1983). *A Guide to Simulation*. New York: Springer-Verlag.
- COCHRAN, W.G. (1963). *Sampling Techniques*, (deuxième édition). New York: John Wiley and Sons.
- DAS, A.K., et TRIPATHI, T.P. (1978). Use of auxiliary information in estimating the finite population variance. *Sankhyā*, 40(C), 139-148.
- DENG, L.Y., et WU, C.F.J. (1987). Estimation of variance of the regression estimator. *Journal of the American Statistical Association*, 82, 568-576.
- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- FULLER, W.A. (1970). Sampling with random stratum boundaries. *Journal of the Royal Statistical Society*, 32, 209-226.
- GARCIA, M.R., et CEBRIAN, A.A. (1996). Repeated substitution method: The ratio estimator for the population variance. *Metrika*, 43, 101-105.
- HIDIROGLOU, M.A., et SÄRNDAAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association, Volume II*, 873-878.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalisation of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- ISAKI, C.T. (1983). Variance estimation using auxiliary information. *Journal of the American Statistical Association*, 78(381), 1117-123.
- MAHAJAN, P.K., et SINGH, S. (1996). On estimation of total in two stage sampling. *Journal of Statistical Research*, 30, 127-131.
- SÄRNDAAL, C.-E. (1996). Efficient estimators with simple variance in unequal probability sampling. *Journal of the American Statistical Association*, 91, 1289-1300.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1989). The weighted residual technique for estimating the variance of the general regression estimator of the finite population total. *Biometrika*, 76(3), 527-537.
- SÄRNDAAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- SAXENA, S.K., NIGAM, A.K., et SHUKLA, N.D. (1995). Variance estimation for combined ratio estimator. *Sankhyā*, 57(B), 85-92.
- SHAH, D.N., et PATEL, P.A. (1996). Asymptotic properties of a generalized regression-type predictor of a finite population variance in probability sampling. *La Revue Canadienne de Statistique*, 24(3), 373-384.
- SINGH, P., et SRIVASTAVA, S.K. (1980). Sampling scheme providing unbiased regression estimators. *Biometrika*, 67, 205-209.
- SINGH, R.K., et SINGH, G. (1984). A class of estimators with estimated optimum values in sample surveys. *Statistics & Probability Letters*, 2, 319-321.
- SINGH, S., et SINGH, S. (1988). Improved estimators of K and B in finite populations. *Journal of the Indian Society of Agricultural Statistics*, 121-126.
- SINGH, S., MANGAT, N.S., et MAHAJAN, P.K. (1995). General class of estimators. *Journal of the Indian Society of Agricultural Statistics*, 47(2), 129-133.
- SRIASTAVA, S.K. (1971). A generalized estimator for the mean of the finite population using multi-auxiliary information. *Journal of the American Statistical Association*, 66, 404-407.
- SRIASTAVA, S.K., et JHAJI, S.K. (1980). A class of estimators using auxiliary information for estimating finite population variance. *Sankhyā* 42(C), 87-96.
- SRIASTAVA, S.K., et JHAJI, H.S. (1981). A class of estimators of the population mean in survey sampling using auxiliary information. *Biometrika*, 68, 341-343.
- SRIASTAVA, S.K., et JHAJI, H.S. (1983). A class of estimators of estimators of the population mean using multi-auxiliary information. *Calcutta Statistical Association Bulletin* 32, 47-56.
- SUKHATME, P.V., et SUKHATME, B.V. (1970). *Sampling Theory of Surveys With Applications*. Iowa: Iowa State University Press.
- SWAIN, A.K.P.C., et MISHRA, G. (1992). Unbiased estimators of finite population variance using auxiliary information. *Metron*, 201-215.
- WU, C.F.J. (1982). Estimation of variance of the ratio estimator. *Biometrika*, 69, 183-189.
- WU, C.F.J. (1985). Variance estimation for combined ratio and combined regression estimators. *Journal of the Royal Statistical Society*, 47(B), 147-154.
- YATES, F., et GRUNDY, P.M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, 15(B), 253-261.

Tableau 3
Comparaison de $\hat{V}_2(\hat{Y}^{\text{RATIO}})$ avec $\hat{V}_1(\hat{Y}^{\text{RATIO}})$ pour une population infinie

n	p	$B[\hat{V}_1(\hat{Y}^{\text{RATIO}})]$	$B[\hat{V}_2(\hat{Y}^{\text{RATIO}})]$	ER	$CIC[\hat{V}_1(\hat{Y}^{\text{RATIO}})]$	$CIC[\hat{V}_2(\hat{Y}^{\text{RATIO}})]$
60	0.1	13.02	10.33	188.7	0.96	0.95
	0.3	8.07	6.35	192.6	0.97	0.95
	0.5	4.33	3.37	195.9	0.96	0.96
	0.7	1.77	1.37	197.9	0.97	0.97
	0.9	0.33	0.26	197.7	0.99	0.98
	0.1	3.27	2.91	123.2	0.94	0.93
80	0.3	2.06	1.84	123.0	0.94	0.94
	0.5	1.13	1.01	122.7	0.95	0.95
	0.7	0.47	0.42	122.0	0.97	0.96
	0.9	0.08	0.08	119.1	0.98	0.97
	0.1	0.76	0.77	106.1	0.94	0.93
	0.3	0.49	0.49	105.8	0.94	0.94
100	0.5	0.27	0.27	105.3	0.95	0.95
	0.7	0.12	0.12	104.4	0.96	0.95
	0.9	0.02	0.02	102.2	0.97	0.95

Tableau 4
Comparaison de $\hat{V}_2(\hat{Y}^{\text{GREG}})$ avec $\hat{V}_1(\hat{Y}^{\text{GREG}})$ pour une population infinie

n	p	$B[\hat{V}_1(\hat{Y}^{\text{GREG}})]$	$B[\hat{V}_2(\hat{Y}^{\text{GREG}})]$	ER	$CIC[\hat{V}_1(\hat{Y}^{\text{GREG}})]$	$CIC[\hat{V}_2(\hat{Y}^{\text{GREG}})]$
60	0.1	10.12	8.42	177.6	0.98	0.95
	0.3	5.06	4.33	161.5	0.97	0.95
	0.5	3.32	2.36	152.5	0.95	0.96
	0.7	0.72	0.38	151.9	0.97	0.95
	0.9	0.13	0.10	147.7	0.99	0.97
	0.1	1.23	1.11	153.9	0.96	0.95
80	0.3	1.03	1.01	143.5	0.98	0.94
	0.5	0.13	0.11	132.8	0.97	0.95
	0.7	0.07	0.06	121.6	0.97	0.95
	0.9	0.02	0.03	117.1	0.96	0.96
	0.1	0.65	0.57	136.1	0.95	0.94
	0.3	0.39	0.32	135.1	0.94	0.94
100	0.5	0.13	0.13	129.6	0.95	0.95
	0.7	0.02	0.02	114.4	0.96	0.95
	0.9	0.01	0.01	112.2	0.97	0.96

Le logiciel SGE mis au point à Statistique Canada peut être adapté afin de fournir de meilleurs estimateurs de la variance de l'estimateur GREG, dont on pourra se servir dans les enquêtes pour lesquelles la variance des variables auxiliaires est connue ou peut être établie.

REMERCIEMENTS

Les auteurs remercient sincèrement le rédacteur associé et les deux examinateurs pour leurs commentaires constructifs et judicieux qui ont permis au manuscrit de trouver sa forme définitive. Ils tiennent aussi à remercier Dr. M.P. Singh pour ses aimables suggestions. Les points de vue et les résultats présentés dans cet article n'engagent que les auteurs et ne reflètent pas nécessairement le point de vue de l'organisme qui les emploie.

7. CONCLUSION

On peut recourir au calage à niveau élevé si on connaît la variance des variables auxiliaires en plus de leur total.

On peut recourir au calage à niveau élevé si on connaît la variance des variables auxiliaires en plus de leur total. L'appliquer en pratique. L'étude empirique s'est effectuée sur FORTRAN-77 au moyen d'un PENTIUM-120.

$$\hat{V}_h(\hat{Y}^{\text{GREG}}) \mid_k \text{ for } h = 1, 2,$$

de la moyenne venant de (1.6) selon un plan d'échantillonnage EASSR. On s'est servi du h -ième estimateur de la variance de l'estimateur de régression pour

Tableau 1
Comparaison de $\hat{V}_2(\hat{Y}^{\text{RATIO}})$ avec $\hat{V}_1(\hat{Y}^{\text{RATIO}})$ pour une population finie

n	$B[\hat{V}_1(\hat{Y}^{\text{RATIO}})]$	$B[\hat{V}_2(\hat{Y}^{\text{RATIO}})]$	ER	$CIC[\hat{V}_1(\hat{Y}^{\text{RATIO}})]$	$CIC[\hat{V}_2(\hat{Y}^{\text{RATIO}})]$
5	-211.33	217.01	166.57	0.93	0.95
6	-141.92	102.00	115.06	0.91	0.92
7	-99.34	58.60	109.23	0.90	0.90

Tableau 2
Comparaison de $\hat{V}_2(\hat{Y}^{\text{GREG}})$ et $\hat{V}_1(\hat{Y}^{\text{GREG}})$ pour une population finie

n	$B[\hat{V}_1(\hat{Y}^{\text{GREG}})]$	$B[\hat{V}_2(\hat{Y}^{\text{GREG}})]$	ER	$CIC[\hat{V}_1(\hat{Y}^{\text{GREG}})]$	$CIC[\hat{V}_2(\hat{Y}^{\text{GREG}})]$
5	-328.49	-194.78	112.04	0.92	0.96
6	-223.92	-136.34	103.02	0.90	0.93
7	-157.88	-94.38	101.21	0.91	0.94

6.4 Populations infinies:

La taille N de ces populations est inconnue. Nous avons produit n paires indépendantes de nombres aléatoires y_i^* et x_i^* (par exemple), $i = 1, 2, \dots, n$, grâce à un sous-programme VNORM avec un coefficient $\text{PHI} = 0.7$, une semence $(y) = 8987878$ et une semence $(x) = 2348789$ conformément aux travaux de Bratley et de ses collaborateurs (1983). Nous avons produit les variables transformées

$$y_i = 3.0 + \sqrt{S_y^2(1 - \rho^2)} y_i^* + \rho S_y x_i^* \quad (6.4.1)$$

$$x_i = 4.0 + S_x x_i^* \quad (6.4.2)$$

avec les valeurs fixes $S_y^2 = 50$ et $S_x^2 = 50$, pour diverses valeurs du coefficient de corrélation ρ . On a calculé l'estimateur

$$\hat{Y}^{\text{RATIO}} | \kappa = \bar{Y} \left(\frac{\bar{X}}{\bar{X}} \right), \text{ avec } \bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i \text{ et } \bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$$

pour le κ -ième échantillon. L'erreur quadratique moyenne empirique de l'estimateur est égale à

$$\text{EQM}(\hat{Y}^{\text{RATIO}}) = \frac{1}{15,000} \sum_{k=1}^{15,000} [\hat{Y}^{\text{RATIO}} | \kappa - \bar{Y}]^2. \quad (6.4.3)$$

On a aussi dérivé, pour le même échantillon, les estimateurs de type ratio de la variance

$$\hat{V}_h(\hat{Y}^{\text{RATIO}}) | \kappa, h = 1, 2,$$

venant de (6.1.1) et de (6.1.2), respectivement, afin d'estimer la variance de l'estimateur de ratio de la moyenne de la population. Le biais du h -ième estimateur de type

$$\hat{Y}^{\text{GREG}} | \kappa = \bar{Y} + \hat{\beta}(\bar{X} - \bar{x})$$

Le processus a été repris avec l'estimateur de régression apparaissant au tableau 3.

Les résultats de ces calculs pour les échantillons de taille $n = 60, 80$ et 100 dont le coefficient de corrélation varie

$$\hat{Y}^{\text{RATIO}} | \kappa \pm 1.96 \sqrt{\hat{V}_h(\hat{Y}^{\text{RATIO}}) | \kappa} \quad (6.4.7)$$

du h -ième estimateur de type ratio de la variance en comparant le nombre de fois où la moyenne réelle de la population \bar{Y} se retrouvait à l'intérieur des limites établies par

$$CIC[\hat{V}_h(\hat{Y}^{\text{RATIO}})] \text{ pour } h = 1, 2$$

la couverture. En utilisant un intervalle de confiance à 95 %, on a établi

$$\text{ER} = \text{EQM} \left\{ \hat{V}_1(\hat{Y}^{\text{RATIO}}) \right\} \times 100 / \text{EQM} \left\{ \hat{V}_2(\hat{Y}^{\text{RATIO}}) \right\} \quad (6.4.6)$$

L'efficacité relative en pourcentage de l'estimateur $\hat{V}_2(\hat{Y}^{\text{RATIO}})$ par rapport à $\hat{V}_1(\hat{Y}^{\text{RATIO}})$ est

$$\text{EQM} \left\{ \hat{V}_h(\hat{Y}^{\text{RATIO}}) \right\} = \frac{1}{15,000} \sum_{k=1}^{15,000} [\hat{V}_h(\hat{Y}^{\text{RATIO}}) | \kappa - \text{EQM}(\hat{Y}^{\text{RATIO}})]^2. \quad (6.4.5)$$

tandis que l'erreur quadratique moyenne est égale à

$$B \left\{ \hat{V}_h(\hat{Y}^{\text{RATIO}}) \right\} = \frac{1}{15,000} \sum_{k=1}^{15,000} \hat{V}_h(\hat{Y}^{\text{RATIO}}) | \kappa - \text{EQM}(\hat{Y}^{\text{RATIO}}) \quad (6.4.4)$$

ratio de la variance correspond à

à l'estimateur

$$\hat{V}_2(\hat{Y}_{\text{GREG}}) = \hat{V}_1(\hat{Y}_{\text{GREG}}) + \psi_3(S_2^2 - s_2^2) \quad (6.2.2)$$

où $\psi_i, i = 1, 2, 3$ a la même signification que celle indiquée précédemment.

Afin de couvrir presque toutes les situations qui reviennent dans la réalité, nous avons envisagé des populations aussi bien finies qu'infinies.

6.3 Populations finies:

Nous avons utilisé une population de $N = 20$ unités venant de Horvitz et Thompson (1952). La variable à l'étude, y_i , correspond au nombre de ménages du i -ième ilot, tandis que la variable auxiliaire connue x représente le nombre de ménages du même ilot, mais estimée visuellement. Tous les échantillons possibles de taille $n = 5$ ont été sélectionnés par BASSR, ce qui a donné

$$\binom{N}{n} = 15,504$$

échantillons. Du k -ième échantillon, on a tiré l'estimateur

$$\hat{Y}_{\text{RATIO}}|_k = \hat{Y} \left(\frac{\bar{X}}{\bar{X}_i} \right), \text{ où } \hat{Y} = \frac{N}{n} \sum_{i=1}^n y_i$$

L'erreur quadratique moyenne empirique de l'estimateur correspond à

$$\text{EQM}(\hat{Y}_{\text{RATIO}}) = \binom{N}{n}^{-1} \sum_{k=1}^K \left[\hat{Y}_{\text{RATIO}}|_k - Y \right]^2. \quad (6.3.1)$$

On a aussi obtenu les estimateurs de type ratio

$$\hat{Y}_h(\hat{Y}_{\text{RATIO}})|_k, h = 1, 2,$$

des équations (6.1.1) et (6.1.2) respectivement, en vue d'estimer la variance de l'estimateur de ratio pour le k -ième échantillon. Le biais du h -ième estimateur de type ratio de la variance est donné par

$$B\left\{\hat{Y}_h(\hat{Y}_{\text{RATIO}})\right\} =$$

$$\binom{N}{n}^{-1} \sum_{k=1}^K \hat{Y}_h(\hat{Y}_{\text{RATIO}}) \left|_k - \text{EQM}(\hat{Y}_{\text{RATIO}}) \right| \quad (6.3.2)$$

tandis que l'erreur quadratique moyenne correspond à

$$\text{EQM}\left\{\hat{Y}_h(\hat{Y}_{\text{RATIO}})\right\} =$$

$$\binom{N}{n}^{-1} \sum_{k=1}^K \left[\hat{Y}_h(\hat{Y}_{\text{RATIO}}) \left|_k - \text{EQM}(\hat{Y}_{\text{RATIO}}) \right| \right]^2. \quad (6.3.3)$$

L'efficacité relative en pourcentage de l'estimateur $\hat{Y}_2(\hat{Y}_{\text{RATIO}})$ par rapport à $\hat{Y}_1(\hat{Y}_{\text{RATIO}})$ s'établit à

$$\text{ER} =$$

$$\frac{\text{EQM}\left\{\hat{Y}_1(\hat{Y}_{\text{RATIO}})\right\} \times 100 / \text{EQM}\left\{\hat{Y}_2(\hat{Y}_{\text{RATIO}})\right\}}{\quad} \quad (6.3.4)$$

En utilisant l'intervalle de confiance à 95 %, on a calculé la couverture de

$$\text{CIC}\left[\hat{Y}_h(\hat{Y}_{\text{RATIO}})\right]$$

pour $h = 1, 2$ pour le h -ième estimateur de type ratio de la variance en comptant combien de fois le total réel de la population Y revenait à l'intérieur des limites définies par

$$\hat{Y}_{\text{RATIO}}|_k \pm t_{n-h-1}(\alpha) \sqrt{\hat{V}_h(\hat{Y}_{\text{RATIO}})|_k}. \quad (6.3.5)$$

On a effectué les mêmes calculs pour tous les échantillons possibles de 6 et de 7 éléments. Les résultats des calculs apparaissent au tableau 1.

La même technique a été employée pour l'estimateur de régression

$$\hat{Y}_{\text{GREG}}|_k = \hat{Y} + \left(\sum_{i=1}^n x_i y_i / \sum_{i=1}^n x_i^2 \right) (X - \bar{X})$$

du total obtenu en (1.6), selon un plan d'échantillonnage BASSR. On s'est servi du h -ième estimateur de la variance de l'estimateur de régression $\hat{Y}_h(\hat{Y}_{\text{GREG}})|_k$ pour $h = 1, 2$, donné respectivement en (6.2.1) et (6.2.2), afin d'établir le

biais, l'efficacité relative (ER) et la CIC. Les résultats se retrouvent au tableau 2. On a noté de surcroît que pour $n = 5$, 0,020 % des estimations de la variance obtenues avec l'estimateur $\hat{Y}_1(\hat{Y}_{\text{GREG}})$ et 0,022 % des estimations obtenues avec l'estimateur $\hat{Y}_2(\hat{Y}_{\text{GREG}})$ étaient négatives. Les populations plus naturelles présentées par Cochran (1963) ou Sukhatme et Sukhatme (1970) donnent des résultats analogues. Dans l'ensemble, les estimateurs de calage du deuxième degré fonctionnent mieux que ceux du premier degré pour les populations finies.

Dans la réalité cependant, il se peut que la variable à l'étude et les variables auxiliaires épousent une distribution d'un certain genre (normale, bêta, gamma, etc.). Pour évaluer la performance des stratégies envisagées en pareilles circonstances, nous avons engendré des populations artificielles et étudié le problème de l'estimation de la moyenne d'une population finie par simulation, de la manière décrite ci-dessous.

où Ω_h désigne des poids soigneusement sélectionnés pour que la fonction de distance chi carré donnée par

$$\sum_{h=1}^L \frac{D_h \bar{\mathcal{O}}_h}{(\Omega_h - D_h)^2} \quad (4.12)$$

ait la valeur minimale, sous réserve d'une équation de calage à niveau élevé définie comme suit

$$\sum_{h=1}^L \Omega_h s_{hx}^2 = V(\bar{x}_{Si}) \quad (4.13)$$

où on suppose que

$$V(\bar{x}_{Si}) = \sum_{h=1}^L W_h^2 \frac{n_h}{(1-f_h)^2} s_{hx}^2$$

est connu et que $s_{hx}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$ constitue un estimateur non biaisé de $S_{hx}^2 = (N_h - 1)^{-1} \sum_{i=1}^{N_h} (X_{hi} - \bar{X}_h)^2$. Cette méthode aboutit à un nouvel estimateur de la variance de l'estimateur de régression combiné représenté par

$$V(\bar{y}_{Si})^{\text{CLR}} = V_{Si}(\bar{y}_{Si})^{\text{GREG}} + B_{Si} \left[V(\bar{x}_{Si}) - V(\bar{x}_{Si}) \right] \quad (4.14)$$

où

$$B_{Si} = \sum_{h=1}^L \frac{n_h}{W_h^2 (1-f_h)} \frac{\bar{\mathcal{O}}_h s_{hx}^2 s_{eh}^2}{\sum_{h=1}^L W_h^2 (1-f_h)} \frac{n_h}{s_{hx}^4}$$

désigne la version améliorée de l'estimateur du coefficient de régression pour l'échantillonnage stratifié et où

$$V(\bar{x}_{Si}) = \sum_{h=1}^L W_h^2 \frac{n_h}{(1-f_h)^2} s_{hx}^2$$

correspond à un estimateur non biaisé de $V(\bar{x}_{Si})$. Si $q_h = 1/\bar{x}_h$ et $\bar{\mathcal{O}}_h = 1/s_{hx}^2$, on peut simplifier l'estimateur (4.14) de façon à obtenir un nouvel estimateur de la variance de l'estimateur de ratio combiné, représenté par

$$V_{Si}(\bar{y}_{Si})^{\text{Ratio}} = \sum_{h=1}^L \frac{n_h}{W_h^2 (1-f_h)} s_{eh}^2 \left(\frac{\bar{x}}{\bar{X}} \right)^2 \left\{ V(\bar{x}_{Si}) \right\} \quad (4.15)$$

ce qui correspond à un estimateur de type ratio proposé par Wu (1985) pour estimer la variance de l'estimateur de ratio combiné, mais pour lequel on recourt aux données supplémentaires sur la variance réelle des variables auxiliaires au moment de l'estimation. On peut constituer plusieurs autres estimateurs en choisissant d'autres poids q_h et $\bar{\mathcal{O}}_h$.

5. PLUS VASTE CATÉGORIE D'ESTIMATEURS

En définissant $u = X / \sum_{i=1}^n d_i x_i$ et $v = V(X_{\text{HT}}) / V(X_{\text{HT}})$, on obtient une catégorie plus vaste d'estimateurs définie comme suit

$$V_{SS}(\bar{y}_{\text{GREG}}) = \left\{ \frac{2}{1} \sum_{j=1}^f \sum_{i=1}^n D_{ij} (d'_i e_j - d_j e_j)^2 \right\} H(u, v) \quad (5.1)$$

où $H(u, v)$ est une fonction paramétrique de u et de v de sorte que $H(1, 1) = 1$ et que certaines conditions de régularité sont satisfaites. Dans ce cas, tous les estimateurs dérivant des fonctions

$$H(u, v) = u^\alpha v^\beta, \quad H(u, v) = \frac{1 + \alpha(u-1)}{1 + \beta(v-1)}, \quad H(u, v) = 1 + \alpha(u-1) + \beta(v-1)$$

et $H(u, v) = \{1 + \alpha(u-1) + \beta(v-1)\}^{-1}$ constituent des cas particuliers de calage à niveau élevé, où α et β sont des paramètres inconnus, intégrés à la fonction $H(u, v)$. En remplaçant ces paramètres par leurs estimateurs convergents respectifs de la catégorie décrite en (5.1), on obtient la même variance asymptotique que Srivastava et Jhaji (1983), Singh et Singh (1984) et Mahajan et Singh (1996). Nous poursuivons présentement l'élargissement de cette analyse à l'échantillonnage à deux phases, dans la foulée des travaux de Hidiroglou and Särndal (1995).

Dans la prochaine partie, nous examinerons la performance de calage à niveau élevé dans le contexte d'une simulation.

6. ÉTUDE DE SIMULATION

En vertu de cette étude, nous avons comparé les estimateurs de la variance de l'estimateur de ratio et de l'estimateur de régression. Pour ne pas semer la confusion, les estimateurs faisant l'objet de la comparaison ont été redéfinis comme suit:

6.1 Estimateur de ratio:

Nous avons comparé les estimateurs de la variance de l'estimateur de ratio

$$V_1(\bar{y}_{\text{Ratio}}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 \left(\frac{\bar{X}}{X} \right)^2 \quad (6.1.1)$$

à l'estimateur

$$V_2(\bar{y}_{\text{Ratio}}) = V_1(\bar{y}_{\text{Ratio}}) \left(\frac{S_x^2}{S_y^2} \right) \quad (6.1.2)$$

6.2 Estimateur de régression:

Nous avons aussi comparé les estimateurs de la variance de l'estimateur de régression

$$V_1(\bar{y}_{\text{GREG}}) = \frac{N^2(1-f)}{n(n-1)} \sum_{i=1}^n e_i^2 + \psi_1(X - \bar{X}) + \psi_2(X - \bar{X})^2 \quad (6.2.1)$$

développement afin de recourir à d'autres fonctions de distance, celles dont parlent Deville et Särndal (1992) par exemple, pour résoudre le problème bidimensionnel à cause de la nature indéfinie des poids D_{ij} . Il reviendra à d'autres de trouver des fonctions de distance qui garantiront la valeur non négative des poids.

4. PLAN D'ÉCHANTILLONNAGE STRATIFIÉ

Supposons que la population se compose de L strates avec N_h éléments dans la h -ième strate desquels on prélève un échantillon aléatoire simple de taille n_h , sans remise. La taille de la population totale est $N = \sum_{h=1}^L N_h$ et celle de l'échantillon, $n = \sum_{h=1}^L n_h$. À la i -ième unité de la h -ième strate sont associées deux valeurs y_{hi} et x_{hi} avec $x_{hi} > 0$ pour covariable. Soit, pour la h -ième strate, les poids $W_h = N_h/N$, $\bar{Y}_h = \bar{Y}_h$, $\bar{X}_h = \bar{X}_h$ la fraction d'échantillonnage $f_h = n_h/N_h$ et $\bar{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$, $\bar{x}_h = \sum_{i=1}^{n_h} x_{hi}/n_h$, en utilisant l'information sur la covariable $\bar{Y} = \sum_{h=1}^L W_h \bar{Y}_h$, $\bar{X} = \sum_{h=1}^L W_h \bar{X}_h$ soit connu. L'idée consiste à estimer \bar{Y} . L'estimateur habituel de la moyenne de la population \bar{Y} est

$$\bar{y}_{SI} = \sum_{h=1}^L W_h \bar{y}_h. \quad (4.1)$$

Envisageons le nouvel estimateur

$$\bar{y}_{SI}^* = \sum_{h=1}^L W_h^* \bar{y}_h \quad (4.2)$$

avec les nouveaux poids W_h^* . Ces derniers sont sélectionnés de telle sorte que la distance de type chi carré exprimée par

$$\sum_{h=1}^L \frac{W_h q_h}{(W_h^* - W_h)^2} \quad (4.3)$$

corresponde au minimum, sous réserve que

$$\sum_{h=1}^L W_h^* \bar{x}_h = \bar{X}. \quad (4.4)$$

La minimisation de (4.3) selon l'équation de calage (4.4) aboutit à l'estimateur de régression de type combiné représenté par

$$\bar{y}_{SI}^* = \sum_{h=1}^L W_h \bar{y}_h + \frac{\sum_{h=1}^L W_h q_h \bar{x}_h}{\sum_{h=1}^L W_h q_h \bar{x}_h^2} \left[\bar{X} - \sum_{h=1}^L W_h \bar{x}_h \right] \quad (4.5)$$

pour le choix optimal des poids indiqué par

$$W_h^* = W_h + \frac{\sum_{h=1}^L W_h q_h \bar{x}_h^2}{\sum_{h=1}^L W_h \bar{x}_h} \left(\bar{X} - \sum_{h=1}^L W_h \bar{x}_h \right) \quad (4.6)$$

Si $q_h = \bar{x}_h^{-1}$, après simplification, l'estimateur (4.5) correspond à l'estimateur de ratio combiné bien connu dont on se sert avec l'échantillonnage stratifié. L'estimateur de la variance bien connu de l'estimateur de régression combiné correspond à

$$V(\bar{y}_{SI}^*) = \sum_{h=1}^L \frac{W_h^2}{W_h^2 (1 - f_h)} \frac{n_h}{s_{eh}^2} \quad (4.7)$$

où

$$s_{e2}^2 = (n_h - 1)^{-1} \sum_{i=1}^{n_h} e_{hi}^2$$

représente la variance de l'échantillon de la h -ième strate alors que $e_{hi} = \bar{y}_{hi} - y_{hi} - b(x_{hi} - \bar{x}_h)$ et $\sum_{h=1}^L W_h q_h \bar{y}_h \bar{x}_h / \sum_{h=1}^L W_h q_h \bar{x}_h$ gardent leur sens habituel. Le calage à bas niveau donne l'estimateur de la variance de l'estimateur de régression combiné suivant

$$V(\bar{y}_{SI}^*)^c = \sum_{h=1}^L \frac{D_h W_h^2}{D_h W_h^*} \frac{n_h}{s_{eh}^2} \quad (4.8)$$

où

$$D_h = \frac{n_h}{W_h^2 (1 - f_h)}$$

et W_h^* apparaissent dans (4.6). Si $q_h = \bar{x}_h^{-1}$, (4.8) donne l'estimateur suivant, après simplification

$$V(\bar{y}_{SI}^*)^{\text{RATIO}} = \left(\frac{\bar{X}}{\bar{x}_{SI}} \right)^2 \sum_{h=1}^L \frac{n_h}{W_h^2 (1 - f_h)} \frac{n_h}{s_{eh}^2} \quad (4.9)$$

qui est un cas particulier d'une catégorie d'estimateurs servant à estimer la variance de l'estimateur de ratio combiné que Wu (1985) exprime comme suit

$$V(\bar{y}_{SI}^*)^W = \left(\frac{\bar{X}}{\bar{x}_{SI}} \right)^W \sum_{h=1}^L \frac{n_h}{W_h^2 (1 - f_h)} \frac{n_h}{s_{eh}^2} \quad (4.10)$$

pour $g = 2$. Saxena et ses collaborateurs (1995) se sont eux aussi penchés sur les propriétés des estimateurs de la variance de l'estimateur de ratio combiné. Pour le calage à niveau élevé, on obtient le nouvel estimateur

$$V_{SI} \left(\bar{Y}_{\text{GREG}} \right) = \sum_{h=1}^L \Omega_{W_h^*} \frac{W_h^2}{s_{eh}^2} \quad (4.11)$$

analyse à la distance D bidimensionnelle de type chi carré (CC) entre deux $n \times n$ grilles créées par les poids Ω_{ij} et pour $i, j = 1, 2, \dots, n$, dans

$$D = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \frac{(\Omega_{ij} - D_{ij})^2}{D_{ij} \bar{D}_{ij}} \quad (3.3)$$

Dans la majorité des cas, $\bar{D}_{ij} = 1$ mais on peut se servir d'autres poids. Nous verrons que l'ajustement de type ratio au moyen de la variance connue des variables auxiliaires forme un cas particulier, pour une valeur précise de \bar{D}_{ij} . La minimisation de (3.3) sous réserve de (3.2) aboutit aux nouveaux poids optimaux représentés par

$$\Omega_{ij} = D_{ij} + \frac{D_{ij} \bar{D}_{ij} (d'_i x'_j - d_j x_j)^2}{\sum_{i=1}^n \sum_{j=1}^n D_{ij} \bar{D}_{ij} (d'_i x'_j - d_j x_j)^4}$$

$$\left[V_{YG}(\bar{X}_{HT}) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d'_i x'_j - d_j x_j)^2 \right] \quad (3.4)$$

pour la valeur optimale du multiplicateur λ de Lagrange

$$\lambda = \frac{V_{YG}(\bar{X}_{HT}) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d'_i x'_j - d_j x_j)^2}{\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} \bar{D}_{ij} (d'_i x'_j - d_j x_j)^4} \quad (3.5)$$

On en trouvera la preuve dans l'annexe. En remplaçant Ω_{ij} dans (3.1) par sa valeur dans (3.4), on obtient l'estimateur de régression

$$\hat{Y}_{SS}(\hat{Y}_{GREG}) = V_{YG}(\hat{Y}_{DS}) + \hat{B}_1 [V_{YG}(\bar{X}_{HT}) - V_{YG}(\bar{X}_{HT})] \quad (3.6)$$

où

$$\hat{B}_1 = \frac{\sum_{i=1}^n \sum_{j=1}^n D_{ij} \bar{D}_{ij} (d'_i x'_j - d_j x_j)^2 (w_i e'_j - w_j e_i)^2}{\sum_{i=1}^n \sum_{j=1}^n D_{ij} \bar{D}_{ij} (d'_i x'_j - d_j x_j)^4}$$

$$= \frac{\hat{\mu}_{22}}{\hat{\mu}_{04}} \quad (\text{par exemple}) \quad (3.7)$$

Cas 3.1: Avec le plan d'échantillonnage EASSR, si $q_i = x'_i$ et $\bar{Q}_{ij} = (d'_i x'_j - d_j x_j)^{-2}$ correspondent aux poids associés au calage à bas niveau et à niveau élevé, respectivement, la stratégie envisagée peut être simplifiée. Avec le calage à niveau élevé, on note les cas que voici. coefficient de régression selon Singh et Singh (1988). donc le considérer comme un meilleur estimateur du exploite le total réel X des variables auxiliaires. On peut retrouver dans (2.1). Le coefficient de régression B_1 se trouve dans $\hat{Y}_{YG}(\bar{X}_{HT}) = \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d'_i x'_j - d_j x_j)^2$ et $V_{YG}(\bar{X}_{HT})$

en $\hat{Y}_{SS}(\hat{Y}_{Ratio}) =$

$$\frac{N^2(1-f)}{n} \times \frac{1}{\sum_{i=1}^n e'_i} \left(\frac{\bar{X}}{S_x^2} \right)^2 \left(\frac{\bar{X}}{S_x^2} \right)^2 \quad (3.8)$$

où $s_x^2 = (n-1)^{-1} \sum_{i=1}^n (x'_i - \bar{x})^2$ représente l'estimateur non biaisé de $S_x^2 = (N-1)^{-1} \sum_{i=1}^N (x'_i - \bar{x})^2$.

Cas 3.2: Si $q_i = 1$ et $\bar{Q}_{ij} = 1 \forall i, j$, on obtient

$$\hat{Y}_{YG}(\hat{Y}_{GREG}) = \frac{N^2(1-f)}{n} \sum_{i=1}^n e'_i + \hat{\Psi}_1(X - \bar{X}) + \hat{\Psi}_2(X - \bar{X})^2 + \hat{\Psi}_3(S_x^2 - s_x^2) \quad (3.9)$$

où $\hat{\Psi}_1$ et $\hat{\Psi}_2$ sont respectivement donnés par (2.9) et (2.10), et

$$\hat{\Psi}_3 = \frac{\sum_{i=1}^n \sum_{j=1}^n (x'_i - x'_j)^4}{N^2(1-f)} \left\{ \sum_{i=1}^n \sum_{j=1}^n (x'_i - x'_j) \left[\frac{(X - \bar{X})(x'_i - x'_j)^2}{\sum_{i=1}^n \sum_{j=1}^n x'_i x'_j} + \frac{(X - \bar{X})(x'_i - x'_j)^2}{2} \right] \right\} \quad (3.10)$$

Sans atténuer la généralité de ce qui précède, on peut dire que les estimateurs de la variance de l'estimateur GREG qui apparaissent en (3.8) et (3.9) ne s'insèrent ni dans l'approche de calage à bas niveau ni dans la catégorie d'estimateurs proposée par Deng et Wu (1987). Ils font partie des analogues des estimateurs proposés par Srivastava et Jhaji (1981) pour estimer la variance de l'estimateur GREG, c'est-à-dire

où $H(\cdot, \cdot)$ est une fonction paramétrique en vertu de laquelle $H(1, 1) = 1$ et qui répond à certaines conditions de régularité définies par ces estimateurs. D'après Srivastava et Jhaji (1981) ainsi que Deng et Wu (1987), vérifier que la catégorie d'estimateurs de (3.11) reste supérieure à celle de (2.11), donc de (2.13), constitue un exercice classique. Une des difficultés que pose (3.1) a trait à la manière d'obtenir des estimations non négatives de la variance par calage. La façon la plus simple revient à optimiser la fonction de distance CC (3.3), sous réserve de la contrainte de calage (3.2) et des conditions $\Omega_{ij} \geq 0 \forall i, j = 1, 2, \dots, n$. Même s'il s'avère difficile de trouver une solution théorique au problème, des techniques de programmation quadratique bien connues peuvent aboutir à des résultats numériques utiles. On ne peut procéder à un simple

$$\psi_2 = \frac{(N - n)}{\sum_n \sum_{j=1}^2 \sum_{i=1}^n (x'_i e'_i - x_j e_j)^2} \left(2N(n-1) \sum_n x_1^2 \right) \quad (2.10)$$

Deng et Wu (1987) ont défini comme suit une classe générale d'estimateurs de la variance de l'estimation

$$V_{YG}^{(F^{DW})} = \frac{N^2(1-f)}{\sum_n \sum_{j=1}^2 e'_j \left(\frac{X}{X'} \right)^g} \quad (2.11)$$

où $e'_j = y'_j - \hat{\beta} x'_j$. La catégorie d'estimateurs (2.11) prend la forme linéaire

$$V_{YG}^{(F^{DW})} = \frac{N^2(1-f)}{\sum_n \sum_{j=1}^2 e'_j} \quad (2.12)$$

$$\left[1 + g \left(\frac{X}{X'} - 1 \right) + \frac{2}{g(g-1)} \left(\frac{X}{X'} - 1 \right)^2 + \dots \right] \quad (2.12)$$

qui ressemble à (2.8). Le calage à bas niveau englobe donc

les estimateurs de la variance des estimateurs du total, c'est-à-dire les techniques d'estimation par régression et par ratio. Fait remarquable, aucune valeur de g ne permet la simplification de (1.6) en l'estimateur du produit de Cochran (1963). C'est pourquoi nous avons laissé de côté l'estimation de la variance de cet estimateur. Pour établir l'efficacité des estimateurs, on examine une classe générale d'estimateurs servant à évaluer la variance de l'estimateur GREG de la manière suggérée par Srivastava (1971), soit

$$V_{YG}^{(F^{GREG})} = \left(\frac{N^2(1-f)}{\sum_n \sum_{j=1}^2 e'_j} \right) H \left(\frac{X}{X'} \right) \quad (2.13)$$

où $H(\cdot)$ est une fonction paramétrique de telle sorte que $H(1) = 1$ et satisfait certaines conditions de régularité. Selon Srivastava (1971), il est facile de constater que les analogues de la catégorie générale d'estimateurs (2.13) donnent la variance minimale de la série d'estimateurs proposée par Deng et Wu (1987) pour l'estimateur de régression et l'estimateur de ratio de Wu (1982). Disons que si on adjoint une fonction quelconque du ratio X/X' à l'estimateur usuel de la variance exprimé par

$$\frac{N^2(1-f)}{\sum_n \sum_{j=1}^2 e'_j} \quad (2.14)$$

la variance asymptotique de l'estimateur résultant ne change pas. En d'autres mots, les estimateurs de la variance de l'estimateur de régression (GREG) du total obtenu par calage à bas niveau gardent une efficacité inférieure ou égale à celle de la série d'estimateurs avancée par Wu (1982) et Deng et Wu (1987). Les poids w_j qui

servent à bâtir l'estimateur de la variance de l'estimateur GREG en (2.1) devient de l'estimation du total de la population et ont donc été baptisés «poids de calage de bas niveau pour l'estimation de la variance». La partie qui suit expose la méthode de calage à niveau élevé où la variance des variables auxiliaires est connue. Plusieurs nouveaux estimateurs constituent des cas particuliers de cette deuxième approche.

3. MEILLEUR ESTIMATEUR DE LA VARIANCE DE L'ESTIMATEUR GREG: CALAGE À NIVEAU ÉLEVÉ

Nous recourrons au calage pour estimer la variance de l'estimateur GREG de (1.6). Les poids D_{ij} établis par Yates et Grundy (1953) pour un estimateur de la variance qui appartiennent en (2.1) sont calés afin que l'estimateur de la variance de la variable auxiliaire donne la variance exacte. Examinons l'estimateur de la variance de l'estimateur GREG

$$V_{SS}^{(F^{GREG})} = \frac{1}{\sum_n \sum_{j=1}^2 \Omega_{ij} (w'_i e'_i - w_j e_j)^2} \quad (3.1)$$

où Ω_{ij} correspond au poids modifié associé à l'expression quadratique de l'estimateur de Yates et Grundy (1953), et est aussi près qu'il se peut de D_{ij} , au sens de la moyenne d'une mesure, sous réserve de l'équation de calage

$$\frac{1}{\sum_n \sum_{j=1}^2 \Omega_{ij} (d'_i x'_i - d_j x_j)^2} = V_{YG}^{(X^{HT})} \quad (3.2)$$

où

$$V_{YG}^{(X^{HT})} = \frac{1}{N} \sum_{j=1}^2 \sum_{i=1}^N (\pi'_i y'_i - \pi_j y_j)(d'_i x'_i - d_j x_j)^2$$

représente la variance connue de l'estimateur du total auxiliaire $X (= \sum_{i=1}^N x_i)$ donné par $X^{HT} = \sum_{i=1}^N d'_i x'_i$. Pour calculer le côté droit de (3.2), on a besoin d'information sur les variables auxiliaires pour chaque unité de la population ou on doit tirer $V_{YG}^{(X^{HT})}$ d'une ancienne enquête ou d'une enquête pilote. Dans certains cas, on possède toute l'information sur les variables auxiliaires, notamment pour le renouvellement des établissements enregistrés, grâce au recensement ou aux dossiers administratifs du Registre des entreprises (RE) ou du Bureau de l'impôt australien (BIA). Das et Tripathi (1978), Singh et Srivastava (1980), Srivastava et Jhaji (1980, 1981), Isaki (1983), Singh et Singh (1988), Swain et Mishra (1992), Shah et Patel (1996) ainsi que Garcia et Cebrian (1996) se sont eux aussi servis de la variance connue des variables auxiliaires. Singh, Mangat et Mahajan (1995) ont examiné diverses catégories d'estimateurs des paramètres inconnus de la population en prenant la variance connue des variables auxiliaires. Fuller (1970) a également émis l'idée d'ajuster les poids D_{ij} en recourant à une méthode similaire à l'estimation par régression. Pour plus de simplicité, nous bornerons notre

Särndal (1989, 1992, 1996) et ses collaborateurs ont analysé l'estimateur de (2.1) à diverses occasions. Cet estimateur en couvre divers autres, ainsi qu'on le verra ci-dessous. Pour plus de simplicité, nous prendrons un plan d'échantillonnage aléatoire simple sans remise (BASSR), c'est-à-dire $\pi_i = \pi_j = n/N$ et $\pi_{ij} = n(n-1)/N(N-1)$. Les cas que voici se présentent.

$$\hat{\phi}^{\text{YG}}(\hat{\phi}^{\text{GREG}}) = \frac{N^2(1-f)}{n} \sum_{i=1}^n e_i' \quad (2.5)$$

où $f = n/N$ et $e_i = y_i - \beta x_i$. Par conséquent, (2.5) représente l'estimateur habituel de la variance de l'estimateur de régression (1.6).

Cas 2.2: Si $q_i = 1/x_i$, l'estimateur (1.6) se simplifie pour donner l'estimateur de ratio du total \hat{Y}_{RATIO} (par exemple). Après simplification, l'estimateur (2.1) aboutit à l'estimateur de la variance de l'estimateur

$$\hat{\phi}^{\text{YG}}(\hat{\phi}^{\text{RATIO}}) = \frac{N^2(1-f)}{n} \sum_{i=1}^n e_i' \left\{ \frac{\hat{X}}{\bar{X}} \right\}_2 \quad (2.6)$$

où

$$e_i = y_i - \left(\frac{\bar{y}}{\bar{x}} \right) x_i \text{ et } \bar{X} = \frac{n}{N} \sum_{i=1}^n x_i.$$

L'estimateur (2.6) est un cas particulier de la catégorie des estimateurs de la variance de l'estimateur de ratio proposée par Wu (1982), soit

$$\hat{\phi}^{\text{YG}}(\hat{\phi}^w) = \frac{N^2(1-f)}{n} \sum_{i=1}^n e_i' \left\{ \frac{\hat{X}}{\bar{X}} \right\}_s \quad (2.7)$$

pour $s = 2$.
Cas 2.3: Si $q_i = 1$ et w_i est donné par (1.5), (2.2) se réalise et (2.1) devient

$$\hat{\phi}^{\text{YG}}(\hat{\phi}^{\text{GREG}}) =$$

$$\frac{N^2(1-f)}{n} \sum_{i=1}^n e_i' (X - \bar{X}) + \psi_1 (X - \bar{X}) + \psi_2 (X - \bar{X})^2 \quad (2.8)$$

où

$$\psi_1 = \frac{(N-n)}{n} \frac{\left(\sum_{i=1}^n x_i' \right)^2}{n(n-1)} - \sum_{i=1}^n \sum_{j=1}^n (e_i' - e_j')(x_i' e_i' - x_j' e_j') \quad (2.9)$$

2. ESTIMATEUR DE LA VARIANCE DE

L'ESTIMATEUR GREG; CALAGE

À BAS NIVEAU

Dans cet article, nous envisagerons l'estimation de la variance de l'estimateur de régression (1.6) à deux niveaux de calage distincts. Le calage à niveau élevé couvre une plus grande diversité d'estimateurs que l'approche retenue par Särndal (1996). En effet, le calage à niveau élevé recourt au total et à la variance connus des variables auxiliaires, tandis que le calage à bas niveau ne fait appel qu'au total connu des variables auxiliaires.

La partie 4 traite du plan d'échantillonnage stratifié. Le poids des strates originales a été calé, ce qui donne des estimateurs de régression et des estimateurs de ratio combinés pour l'échantillonnage par stratification. On verra que les estimateurs de la variance des estimateurs de régression et de ratio combinés proposés par Wu (1985) constituent des cas spéciaux de calage à bas niveau. Le calage à niveau élevé peut s'appliquer à une plus grande variété d'estimateurs.

$$\hat{\phi}^{\text{YG}}(\hat{\phi}^{\text{DS}}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (w_i e_i' - w_j e_j')^2 \quad (2.1)$$

où $D_{ij} = (\pi_i \pi_j - \pi_{ij}) / \pi_{ij}$, $i \neq j$ et $e_i = y_i - \beta x_i$ gardent leur sens habituel. On peut reformuler facilement cet estimateur comme suit

$$\hat{\phi}^{\text{YG}}(\hat{\phi}^{\text{DS}}) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i e_i' - d_j e_j')^2 +$$

$$\psi_1 \left(X - \bar{X} - \sum_{i=1}^n d_i x_i' \right) + \psi_2 \left(X - \bar{X} - \sum_{i=1}^n d_i x_i' \right)^2 \quad (2.2)$$

où

$$\psi_1 = \frac{\sum_{i=1}^n d_i' q_i' x_i'^2}{1}$$

et

$$\sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i' e_i' - d_j' e_j') (d_i' q_i' x_i' e_i' - d_j' q_j' x_j' e_j') \quad (2.3)$$

$$\psi_2 = \frac{2 \left(\sum_{i=1}^n d_i' q_i' x_i'^2 \right)^2}{1 \sum_{i=1}^n \sum_{j=1}^n D_{ij} (d_i' q_i' x_i' e_i' - d_j' q_j' x_j' e_j')^2} \quad (2.4)$$

Estimation de la variance de l'estimateur général de régression : approche de calage à niveau élevé

SARJINDER SINGH, STEPHEN HORN et FRANK YU¹

RÉSUMÉ

L'analyse qui suit porte sur le problème qui consiste à estimer la variance de l'estimateur général de régression linéaire. On a montré que l'approche de calage à bas niveau adoptée par Särndal (1996) est moins ou aussi efficace que celle d'une catégorie d'estimateurs avancée par Deng et Wu (1987). On a aussi proposé une approche par calage à un niveau plus élevé. Les auteurs montrent que cette dernière constitue une amélioration par rapport à l'originale. Plusieurs estimateurs correspondent à des cas particuliers de la nouvelle approche. On a notamment émis l'idée d'obtenir une estimation non négative de la variance de l'estimateur GREG. Les résultats ont été appliqués à un plan d'échantillonnage aléatoire stratifié. On a aussi entrepris une étude empirique afin de jauger l'efficacité des stratégies envisagées. Le logiciel de statistique SGE bien connu, élaboré par Statistique Canada, peut être perfectionné en vue de fournir une estimation plus précise de la variance de l'estimateur GREG par calage à niveau élevé, dans certaines circonstances examinées plus bas.

MOTS CLÉS : Calage; estimation de la variance; données auxiliaires; estimateurs de ratio et de régression; approche assistée d'un modèle.

1. INTRODUCTION

Le statisticien s'intéresse souvent à la précision des estimations d'enquêtes. L'estimateur de la moyenne ou du total d'une population le plus couramment utilisé est l'estimateur de régression généralisé (GREG). Examinons le cas le plus simple de l'estimateur GREG quand on ne possède de l'information que sur une variable auxiliaire. Soit une population $\Omega = \{1, 2, \dots, N\}$, de laquelle on prélève un échantillon probabiliste $s (s \subset \Omega)$ selon un plan d'échantillonnage $p(\cdot)$ établi. On présume que les probabilités d'inclusion $\pi_i = Pr(i \in s)$ et $\pi_{ij} = Pr(i \text{ et } j \in s)$ sont positives et connues. Soit y_i , la valeur de la variable y à laquelle on s'intéresse pour le i -ième élément de la population, également associé à une variable auxiliaire x_i . Pour les éléments $i \in s$, on observe (y_i, x_i) . On suppose que le total de la variable x , pour la population, $X = \sum_{i=1}^N x_i$, est connu de façon précise. L'objectif consiste à estimer le total de la population $Y = \sum_{i=1}^N y_i$. Deville et Särndal (1992) calent le total connu de la population x afin de modifier les poids de base du plan d'échantillonnage $d_i = 1/\pi_i$, de l'estimateur de Horvitz-Thompson (1952)

$$\hat{Y}_{HT} = \sum_{i=1}^n \frac{y_i}{x_i} = \sum_{i=1}^n d_i y_i. \quad (1.1)$$

Deville et Särndal (1992) avaient proposé un nouvel estimateur

$$\hat{Y}_{DS} = \sum_{i=1}^n w_i y_i \quad (1.2)$$

dont les poids w_i se situaient aussi près que possible de d_i , dans le sens d'une mesure moyenne, sans qu'en souffre l'équation de calage

$$\sum_{i=1}^n w_i x_i = X. \quad (1.3)$$

Un cas simple, examiné par Deville et Särndal (1992), concerne la minimisation de la fonction de distance de type chi carré représentée par

$$\sum_{i=1}^n \frac{d_i q_i}{(w_i - d_i)^2} \quad (1.4)$$

où q_i désigne les poids appropriés. Dans la plupart des situations, $q_i = 1$. La forme de l'estimateur dépend du poids q_i choisi. En minimisant (1.4) sous réserve de l'équation de calage (1.3), on obtient les poids

$$w_i = d_i + \frac{\sum_{i=1}^n d_i q_i x_i}{d_i q_i x_i} \left(X - \sum_{i=1}^n d_i x_i \right). \quad (1.5)$$

Il suffit de remplacer la valeur de w_i dans (1.2) par celle de (1.5) pour parvenir à l'estimateur de régression classique du total

$$\hat{Y}_{DS} = \sum_{i=1}^n d_i y_i + \frac{\sum_{i=1}^n d_i q_i x_i^2}{\sum_{i=1}^n d_i q_i x_i} \left(X - \sum_{i=1}^n d_i x_i \right). \quad (1.6)$$

¹ Sarjinder Singh, Research Officer, Stephen Horn, Senior Research Officer et Frank Yu, Director, Methodology Division, The Australian Bureau of Statistics, P.O. Box 10, Belconnen, ACT 2616, Australie.

- JESSEN, R. J. (1942). Statistical investigation of a sample survey for obtaining farm facts. *Iowa Agricultural Experiment Station Research Bulletin*, 304, 54-59.
- JONES, R.G. (1980). Best linear unbiased estimators for repeated surveys. *Journal of the Royal Statistical Society, Series B*, 42, 221-226.
- KUMAR, S., et LEE, H. (1983). Évaluation de l'application d'estimateurs composites à l'enquête sur la population active au Canada. *Techniques d'enquête*, 9, 196-221.
- LENT, J., MILLER, S.M., et CANTWELL, P.J. (1996). Effect of composite weights on some estimates from the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 130-139.
- ODELL, P.L., et LEWIS, T.O. (1971). Best linear recursive estimation. *Journal of the American Statistical Association*, 66, 893-896.
- PATTERSON, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society, Series B*, 12, 241-255.
- PFEFFERMANN, D. (1991). Estimation and seasonal adjustment of population means using data from repeated surveys. *Journal of Business and Economic Statistics*, 9, 163-175.
- RAO, J.N.K., et GRAHAM, J.E. (1964). Rotation designs for sampling on repeated occasions. *Journal of the American Statistical Association*, 59, 492-509.
- SALLAS, W.M., et HARVILLE, D.A. (1981). Best linear recursive estimation for mixed linear models. *Journal of the American Statistical Association*, 76, 860-869.
- SCOTT, A.J., SMITH, T.M.F., et JONES, R.G. (1977). The application of time series methods to the analysis of repeated surveys. *Revue Internationale de Statistique*, 45, 13-28.
- SINGH, A. C. (1996). Combining information in survey sampling by modified regression. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 120-129.
- TILLER, R. (1989). A Kalman filter approach to labor force estimation using survey data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 16-25.
- WOLTER, K.M. (1979). Composite estimation in finite populations. *Journal of the American Statistical Association*, 74, 604-613.
- YANSANEH, I.S. (1992). Least Squares Estimation for Repeated Surveys. Thèse de doctorat, non-publié. Département of Statistics, Iowa State University, Ames, Iowa.
- YANSANEH, I.S. (1997). Recursive regression estimation in the presence of time-in-sample effects. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 164-169.
- ADAM, A., et FULLER, W.A. (1992). Covariance estimators for the Current Population Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 586-591.
- BAILLAR, B.A. (1975). The effects of rotation group bias on estimates from panel surveys. *Journal of the American Statistical Association*, 70, 23-29.
- BELL, W.R., et HILLMER, S.C. (1990). Estimation dans les enquêtes à passages répétés au moyen de séries chronologiques. *Techniques d'enquête*, 16, 205-227.
- BINDER, D.A., et DICK, J.P. (1989). Enquêtes répétées – Modélisation et estimation. *Techniques d'enquête*, 15, 31-48.
- BREAU, P., et ERNST, L. (1983). Alternatives estimators to the current composite estimator. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 397-402.
- COCHRAN, W.G. (1942). Sampling theory when the sampling units are of unequal sizes. *Journal of the American Statistical Association*, 37, 199-212.
- DUNCAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.
- FULLER, W.A. (1990). Analyse d'enquêtes à passages répétés. *Techniques d'enquête*, 16, 177-190.
- FULLER, W.A., ADAM, A., et YANSANEH, I.S. (1993). Estimateurs pour des enquêtes longitudinales avec application à l'enquête. *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales*, Statistique Canada, Ottawa, Canada, 349-366.

BIBLIOGRAPHIE

$$\sum_{j=1}^J \gamma_j \hat{\theta}_{t-j}(t) = \hat{\theta}_t(t) + \sum_{j=1}^J \gamma_j \hat{\theta}_{t-j}(t) \quad (A3)$$

$$= \sum_{s=1}^s \sum_{h=1}^h \sum_{i=1}^i c_{ih} \gamma_{i,h} + \sum_{m=1}^m \sum_{s=1}^s \sum_{h=1}^h f_{ih(t-j)} \gamma_{i,h}$$

$$= \sum_{s=1}^s \left[c_{ih} + \sum_{m=1}^m \sum_{h=1}^h \sum_{i=1}^i \gamma_{ij} f_{ih(t-j)} \right] \gamma_{i,h}$$

où $c_{ih}, i = 1, 2, \dots, s$, sont les coefficients de $\gamma_{i,t}$ dans $\hat{\theta}_{t-j}(t)$, $j = 1, \dots, m$, sont les coefficients de $\gamma_{i,t}$ dans $\hat{\theta}_{t-j}(t)$, $j = 1, \dots, m$, respectivement. Par conséquent, $\sum_{i=1}^s c_{ih} = 1$, et $\sum_{i=1}^s \gamma_{ij} f_{ih(t-j)} = 1$. Ainsi dans la combinaison linéaire (A3), la somme des coefficients des observations $\gamma_{i,t}, i = 1, 2, \dots, s$, au temps t est égale à un. Par conséquent, au moins un des coefficients est supérieur ou égal à s^{-1} . Donc, $\text{Var}\{\sum_{j=0}^m \gamma_j \hat{\theta}_{t-j}(t)\} \geq s^{-2} \gamma_{mm}^2$, et on en conclut que $\sum_{j=0}^m \gamma_j \hat{\theta}_{t-j}(t)$ a une valeur positive définie.

Preuve. Pour commencer, on montre que la variance de l'estimateur des moindres carrés de $\theta^c - \theta^{c-1}$ (designant la période courante) converge à mesure que le nombre de périodes augmente, en reprenant l'argumentation du premier lemme. On peut aussi se servir d'une argumentation semblable à celle du deuxième lemme pour montrer que les variances des estimateurs des moindres carrés des paramètres $\theta^{l-m}, \theta^{l-m+1}, \dots, \theta^{l-1}$ convergent toutes lorsque le nombre de périodes augmente.

Preuve du théorème. Puisque $\sum_{i=0}^{l(m)} \theta^{i-j}$ est une sous-matrice de la matrice de covariance $\sum_{i=0}^{l(m+1)} \theta^{i-j}$ des estimateurs des moindres carrés du jeu complet de paramètres $\theta^{l-m}, \theta^{l-m+1}, \dots, \theta^{l-1}, \theta^l$, au temps t , il suffit de montrer que $\sum_{i=0}^{l(m+1)} \theta^{i-j}$ converge vers une matrice de valeur positive définie quand $t \rightarrow \infty$. Des lemmes 1 et 2, il s'ensuit que chaque élément diagonal de $\sum_{i=0}^{l(m+1)} \theta^{i-j}$ converge vers un nombre positif à mesure que $t \rightarrow \infty$. Du lemme 3, il ressort que la variance de l'estimateur des moindres carrés de chaque paramètre $\theta^{l-m}, \theta^{l-m+1}, \dots, \theta^{l-1}, \theta^l$ converge vers un nombre positif quand $t \rightarrow \infty$. Il en découle que pour chaque valeur j , $1 \leq j \leq m$, la covariance entre les estimateurs des moindres carrés de θ^{l-j} et de $\theta^{l-j'}$ converge lorsque $t \rightarrow \infty$ tandis que la matrice de covariance $\sum_{i=0}^{l(m+1)} \theta^{i-j}$ converge quand $t \rightarrow \infty$.

Prouvons maintenant que la matrice de covariance servant de limite à une valeur positive définie. Soit $\lim_{t \rightarrow \infty} \sum_{i=0}^{l(m)} \theta^{i-j} = \sum_{i=0}^{l(m)} \theta^{i-j}$. Il suffit de montrer que la variance de toute combinaison linéaire non triviale des estimateurs des moindres carrés $\theta^{l-j}(t), j = 1, 2, \dots, m$ a comme borne inférieure une quantité positive. Soit v_{mm} la limite inférieure de chaque combinaison linéaire des observations, l'un des coefficients étant égal à un. La limite est positive du fait qu'on suppose que les éléments de V^{-1} sont bornés.

Chaque estimateur du paramètre $\theta^{l-j}, j = 0, 1, \dots, m$ correspond à une combinaison linéaire de toutes les observations, de sorte que la somme des coefficients des observations des s bandes au temps $t - j$ donne un et que la somme des coefficients des observations des s bandes à tout autre moment est égale à zéro. Cette condition est essentielle pour que l'estimateur ne présente pas de biais au temps $t - j$. Pour que la somme des coefficients des s observations au temps $t - j$ soit égale à un, au moins un des coefficients doit être supérieur ou égal à s^{-1} . La variance minimale d'une combinaison linéaire dont le premier coefficient doit être égal à s^{-1} est $s^{-2} v_{mm}$. Donc, pour $j = 0, 1, \dots, m$, $\text{Var}\{\hat{\theta}^{l-j}(t)\} \geq s^{-2} v_{mm}$. Envisageons maintenant, une combinaison linéaire non triviale arbitraire des estimateurs récursifs des moindres carrés $\hat{\theta}^{l-j}(t), j = 0, 1, \dots, m$, donnée par $\sum_{j=0}^m \gamma_j \hat{\theta}^{l-j}(t)$, où, sans réduire la généralité $\gamma_0 = 1$. On peut alors écrire cette combinaison linéaire sous la forme:

Lemme 2. Supposons que les hypothèses du théorème tiennent. La variance de l'estimateur des moindres carrés de chaque paramètre $\theta^{l-m}, \theta^{l-m+1}, \dots, \theta^{l-1}$, reposant sur les données recueillies jusqu'au temps t converge vers un nombre positif à mesure que t augmente.

Preuve. Premièrement, supposons qu'à un temps fixe τ , on dispose d'au moins m périodes d'observations avant et après τ . Définissons une transformation de la forme qui suit pour les observations de chacune des s bandes au temps τ : $u_{i\tau} = y_{i\tau} - \sum_{j=-m}^m b_{k(i,\tau),j} y_{i,\tau-j}$, où $b_{k(i,\tau),0} = 0$ et $u_{i\tau}$ ne présentent pas de corrélations avec les observations qui précèdent et suivent $y_{i\tau}$ dans la i -ième bande. Soit la variance de $u_{i\tau}, i = 1, 2, \dots, s$. Ces variances comportent une limite inférieure, établie par hypothèse. Comme auparavant, on conclut qu'il existe une limite positive inférieure aux éléments diagonaux de la matrice de covariance du vecteur des estimateurs récursifs des moindres carrés.

Supposons maintenant qu'au temps t , la séquence d'estimation débute avec le vecteur des estimateurs récursifs des moindres carrés $\hat{\theta}^{l-1(m)} = (\hat{\theta}^{l-m}, \dots, \hat{\theta}^{l-1})'$ reposant sur les données des m périodes antérieures et le vecteur des observations transformées $z_t' = (z_{1t}, \dots, z_{st})'$. Le modèle linéaire des données au temps t correspond donc à (7), où l'on remplace c par t . Le nombre de dimensions du vecteur de données z_t' est fixe. Par conséquent, la matrice de covariance du MELSB pour le vecteur des paramètres $\theta^{l-1(m)} = (\theta^{l-m}, \dots, \theta^{l-1})'$ est $(W'V^{-1}W)^{-1}$. Pour faciliter les calculs, on exprime W comme (I_{m+1}', M') , où I_{m+1} représente la matrice unité de degré $m+1$, et M est une $(s-1) \times (m+1)$ matrice, constante dans le temps. On obtient donc

$$\sum_{i=0}^{l(m+1)} (\Omega_{l-1}^{l-1(m)} + M' \Omega_{l-1}^{00} M^{-1}) = \Omega_{l-1}^{l-1(m)} - \Omega_{l-1}^{l-1(m)} M D_{l-1}' M \Omega_{l-1}^{l-1(m)} \quad (A2)$$

où $\Omega_{l-1}^{l-1(m+1)} = \text{blockdiag}\{\sum_{i=0}^{l-1(m)} \sigma_1^2, \sigma_2^2, \dots, \sigma_s^2\}$, et $D_{l-1} = \Omega_{l-1}^{00} + M \Omega_{l-1}^{l-1(m+1)} M'$. Puisque le deuxième terme à droite de (A2) a une valeur positive définie, on en conclut que les m premiers termes diagonaux de $\sum_{i=0}^{l(m+1)} \theta^{i-j}$ sont inférieurs ou égaux aux éléments diagonaux originaux de $\sum_{i=0}^{l(m)} \theta^{i-j}$. Bref, la variance des estimateurs de $\theta^{l-m}, \dots, \theta^{l-2}, \theta^{l-1}$ n'augmente pas avec t . Puisque les variances en question ont une limite inférieure positive, on en conclut que les variances des estimateurs de $\theta^{l-m}, \dots, \theta^{l-2}, \theta^{l-1}$ convergent vers un nombre positif à mesure que t augmente.

Lemme 3. Supposons que les hypothèses du théorème tiennent. Dans ce cas, la variance de l'estimateur des moindres carrés de chaque paramètre $\theta^{l-m}, \theta^{l-m+1}, \dots, \theta^{l-1}$, jusqu'au temps t , converge vers un nombre positif à mesure qu'augmente t .

3. Le plan de renouvellement intermédiaire 4-8-4 s'avère moins efficace que les plans de renouvellement continu pour les changements survenant sur de courtes périodes, mais il demeure préférable pour le niveau courant, les moyennes couvrant de longues périodes et la variation des moyennes relatives à de longues périodes.
4. L'estimateur composite de la CPS est comparable à l'ERR pour ce qui est d'estimer le changement d'une et de 12 périodes à l'égard des chômeurs. La méthode d'estimation par régression récursive y est cependant supérieure pour les autres mesures du changement.
5. L'ERR est plus efficace pour estimer le changement de niveau avec les décalages pour lesquels l'estimateur composite de la CPS n'a pas été conçu, par exemple ceux de quatre à six mois.

REMERCIEMENTS

Les auteurs remercient l'examinateur et John Bltinge pour leurs précieux commentaires concernant les versions antérieures du présent article, dont la réalisation a été financée en partie par le U.S. Bureau of the Census (Joint Statistical Agreement 91-21), ainsi que par le National Agricultural Statistics Service et le U.S. Bureau of the Census (Cooperative Agreement 43-3ABU-80088). Les opinions exprimées dans le présent document n'engagent que leurs auteurs et ne correspondent pas nécessairement aux politiques du Bureau of the Census, du Bureau of Labor Statistics ou du National Agricultural Statistics Service.

ANNEXE

Lemme 1. Supposons que les hypothèses du théorème tiennent. La variance de l'estimateur au niveau courant θ_c converge alors vers un nombre positif, à mesure que le nombre de périodes augmente.

Preuve. Si on connaît les moyennes $\theta_{c-1}, \theta_{c-2}, \dots, \theta_{c-m}$, $g_{1c}, i = 1, 2, \dots, s$ sont des estimateurs non biaisés de θ_c , où $g_{1c} = y_{1c}, g_{2c} = y_{2c} - b_{21}(y_{2c-1} - \theta_{c-1}), \dots$; et $g_{sc} = y_{sc} - \sum_{j=1}^m b_{sj}(y_{sj,c-j} - \theta_{c-j})$. Par ailleurs, $g_{1c}, i = 1, 2, \dots, s$ sont indépendants et leur variance est égale à $\sigma_i^2, i = 1, 2, \dots, s$. On peut formuler le modèle linéaire:

$$\mathbf{g} = \mathbf{J}^s \theta_c + \mathbf{e} \quad (\text{A1})$$

où $\mathbf{g} = (g_{1c}, g_{2c}, \dots, g_{sc})'$, \mathbf{J}^s correspond au $s \times 1$ vecteur colonne de valeurs un et \mathbf{e} , au $s \times 1$ vecteur des erreurs où $E(\mathbf{e}) = 0$, et $E(\mathbf{ee}) = \mathbf{V}^s = \text{Diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_s^2)$. Le MELSB de θ_c du (A1) a donc pour variance $(\sum_{i=1}^s \sigma_i^2)^{-1}$. Par hypothèse, les variances $\sigma_i^2, i = 1, 2, \dots, s$ ont une limite inférieure et la quantité $(\sum_{i=1}^s \sigma_i^2)^{-1}$ constitue une limite positive inférieure de la variance de l'estimateur θ_c (voir le lemme 4.2.3 de Yansaneh (1992)). La variance de l'estimateur de θ_c n'augmente pas avec le nombre d'observations, si bien qu'elle tend vers un nombre positif.

Lorsqu'on veut estimer les moyennes des personnes occupées pour 12 périodes avec le plan 4-8-4, on se rend compte que l'estimateur composite de la CPS est environ 13 % moins efficace que l'estimateur des moindres carrés. La perte d'efficacité s'établit à environ 53 % lorsqu'il s'agit d'estimer le changement, ainsi qu'on peut le voir en comparant la deuxième et la troisième colonne du tableau 2. Pour les chômeurs et le plan 4-8-4, la précision ne s'accroît que de façon minime quand on préfère l'estimateur des moindres carrés à l'estimateur composite de la CPS, ainsi qu'en témoignent les sixième et septième colonnes du tableau 2.

Le plan 4-8-4 s'avère nettement supérieur aux deux plans à renouvellement continu pour ce qui est d'estimer le changement de 12 périodes, la moyenne de 12 périodes et l'écart des moyennes de 12 périodes, pour les deux caractéristiques. En général, les plans à renouvellement continu donnent de meilleurs résultats avec les changements d'une période à l'autre, sur de brèves périodes.

6.3.3 Cohérence interne

Dans notre analyse, nous avons conçu le meilleur estimateur pour les personnes occupées en ne nous servant que des données passées sur ces personnes et les chômeurs. Rien n'interdit l'inclusion des données antérieures sur les personnes occupées et les chômeurs au moment de construire les estimateurs. Selon Fuller et ses coll. (1993) cependant, les corrélations croisées estimées ne dépassent pas 0,10, signe que pareille inclusion n'améliorerait pas grand-chose.

Fuller (1990) a suggéré une méthode pour obtenir des estimations sur des caractères multiples gardant une cohérence interne. Il procède pour cela à des estimations sur les personnes occupées, les chômeurs et les personnes hors de la population active, puis s'en sert comme valeurs témoins lors d'une régression servant à établir les poids applicables aux observations courantes. Ces poids peuvent alors être utilisés pour obtenir des estimations internes cohérentes de n'importe quel paramètre digne d'intérêt. Cette méthode d'estimation, y compris l'estimation des sous-groupes de la population active, devrait être appliquée à la CPS en 1998 (voir Lent, Miller et Cantwell 1996).

6.4 Conclusions

Voici, en résumé, les principales conclusions qui émergent des calculs de la variance présentés ici.

1. D'autres méthodes d'estimation donnent une plus faible variance du niveau courant que l'estimateur composite de la CPS, pour tous les plans de renouvellement et toutes les caractéristiques à l'étude.
2. Le gain de précision des autres estimateurs, par rapport à l'estimateur composite de la CPS, augmente avec le décalage et atteint un maximum quand le décalage correspond au chevauchement minimal, lorsqu'on estime le changement avec le plan de renouvellement 4-8-4.

utilisé dans l'Enquête sur la population active canadienne. À chaque période, l'échantillon comporte six groupes de renouvellement: un interviewé pour la première fois et ainsi de suite jusqu'au dernier, qui fait partie de l'échantillon depuis six mois. Chaque groupe reste donc dans l'échantillon six mois de suite avant d'en être retranché. Kumar et Lee (1983) donnent plus de détails sur le plan d'échantillonnage de l'Enquête sur la population active. L'autre plan compte 8 groupes de renouvellement à chaque période et chaque groupe demeure 8 périodes dans l'échantillon, avant d'être abandonné.

Nous avons comparé la performance des divers plans de renouvellement au moyen du MELSB du niveau courant pour 36 périodes. À 36 périodes, l'estimateur des moindres carrés est presque aussi efficace que l'ERR. Pour tous les plans de renouvellement examinés, l'usage du meilleur estimateur au lieu de l'estimateur composite de la CPS

La précision des estimateurs du changement de niveau par rapport à celle de l'estimateur composite de la CPS dépend du plan de renouvellement. Ainsi, au tableau 2, on constate que le plan 4-8-4 accroît quelque peu la précision et que ce gain augmente avec le décalage. Dans le cas des personnes occupées, la variance de l'estimateur de l'estimateur des moindres carrés est égale à 85 % de la variance de l'estimateur composite. Ainsi, au tableau 2, le gain le plus élevé concerne les personnes occupées avec le plan de renouvellement 4-8-4, puisque la variance du meilleur estimateur du niveau courant ne correspond qu'à 76 % de la variance de l'estimateur composite.

et (3) du tableau 2).

Tableau 2

Variance des estimateurs et plans de renouvellement; la variance de l'estimateur de base du niveau courant de chaque plan est égale à un

Personnes occupées		Chômeurs	
Paramètre	(1)	Comp. CPS	(2)
	(2)	(3)	(4)
	(4-8-4)	(8 cont.)	(6 cont.)
Niveau courant	0,862	0,653	0,761
Ecart 1	0,511	0,432	0,395
Ecart 2	0,813	0,604	0,559
Ecart 3	1,065	0,710	0,669
Ecart 4	1,279	0,783	0,731
Ecart 5	1,363	0,828	0,782
Ecart 6	1,390	0,854	0,828
Ecart 7	1,388	0,863	0,874
Ecart 8	1,353	0,858	0,828
Ecart 9	1,255	0,830	0,960
Ecart 10	1,154	0,803	0,993
Ecart 11	1,061	0,779	1,021
Ecart 12	0,992	0,758	1,046
Moy. 12 périodes	0,369	0,326	0,440
Ecart Moy. 12 périodes	0,248	0,162	0,365
	0,403	0,273	0,262
	0,255	0,249	0,301
	1,593	1,564	1,696
	1,614	1,578	1,688
	1,641	1,595	1,678
	1,671	1,614	1,663
	1,701	1,934	1,642
	1,710	1,636	1,612
	1,708	1,628	1,577
	1,691	1,606	1,533
	1,645	1,562	1,473
	1,528	1,473	1,372
	1,361	1,338	1,250
	1,070	1,073	1,003
	0,947	0,918	0,944
	0,759	0,938	

Tableau 1
Variance des autres estimateurs par rapport à celle de l'estimateur de base du niveau courant

Personnes occupées					Chômeurs				
Paramètre	Comp. CPS	MELSB 12	MELSB 16	Est. de régression réursive	(1)	Comp. CPS	MELSB 12	MELSB 16	Est. de régression réursive
Niveau courant	0,862	0,704	0,672	0,650	0,947	0,924	0,918	0,918	0,918
Ecart 1	0,511	0,457	0,437	0,432	1,070	1,077	1,073	1,073	1,073
Ecart 2	0,813	0,646	0,613	0,604	1,361	1,345	1,338	1,338	1,338
Ecart 3	1,065	0,763	0,724	0,711	1,528	1,481	1,473	1,473	1,473
Ecart 4	1,279	0,830	0,800	0,784	1,645	1,569	1,563	1,562	1,562
Ecart 5	1,363	0,880	0,847	0,829	1,691	1,614	1,607	1,606	1,606
Ecart 6	1,390	0,910	0,873	0,855	1,708	1,637	1,628	1,628	1,628
Ecart 7	1,388	0,930	0,884	0,865	1,710	1,646	1,637	1,636	1,636
Ecart 8	1,353	0,932	0,884	0,860	1,701	1,645	1,635	1,634	1,634
Ecart 9	1,255	0,912	0,854	0,832	1,671	1,624	1,614	1,614	1,614
Ecart 10	1,154	0,895	0,824	0,806	1,641	1,606	1,595	1,595	1,595
Ecart 11	1,061	0,883	0,795	0,782	1,614	1,590	1,578	1,578	1,578
Ecart 12	0,992	0,883	0,767	0,761	1,593	1,577	1,563	1,563	1,563

ceux de l'estimateur composite de la CPS, pour tous les paramètres, sauf le changement d'une période pour les chômeurs. Rappelons que l'estimateur du changement ne donne pas le MELSB parce qu'il s'agit de l'écart entre les estimateurs construits au temps t et au temps $t - 1$. Le MELSB ne constitue donc le meilleur estimateur du niveau courant que lorsqu'on se sert de la somme de données indiquée. L'écart entre la variance de l'estimateur composite pour le changement d'une période et celle du MELSB de 12 périodes pour le changement d'une période est inférieur à un pour cent, dans le cas des chômeurs. La meilleure méthode d'estimation linéaire sans biais pour les personnes occupées gagne 22 % de précision sur l'estimateur composite de la CPS quand elle repose sur 12 périodes, 28 % quand elle repose sur 16 périodes et 30 % sur 24 périodes. Le gain de précision s'établit à 33 % pour l'ERR. En ce qui concerne les chômeurs, la précision de l'estimation s'améliore respectivement de 2 %, de 3 % et de 3 %. Ces résultats traduisent la nature des fonctions d'autocorrélation des caractéristiques. L'autocorrélation s'affaiblit beaucoup plus vite avec les chômeurs qu'avec les personnes occupées.

Le tableau 2 présente la variance de la comparaison des nouveaux estimateurs selon divers plans de renouvellement. Cette variance se rapporte toujours à celle de l'estimateur de base du niveau courant pour le plan en question. On a comparé l'efficacité des estimateurs à l'étude pour le niveau courant, le changement de niveau et le niveau moyen pour plusieurs périodes avec le plan de renouvellement intermittent 4-8-4 et deux plans de renouvellement continu, soit ceux à 6 et à 8 groupes. Le premier correspond à celui

6.3.2 Comparaison des estimateurs et des plans de renouvellement

Si on néglige le changement d'une période pour les personnes occupées, l'estimation du changement s'améliore quand on abandonne l'estimateur composite de la CPS pour les autres estimateurs. Le gain de précision augmente avec le nombre de périodes, pour atteindre un maximum au bout de cinq périodes, pour les deux caractéristiques. Ensuite, le gain diminue légèrement. Le gain maximal d'efficacité de l'ERR à l'égard de l'estimation du changement s'établit pour sa part à 64 % avec les personnes occupées et à 5 % avec les chômeurs.

6. APPLICATION À LA CURENT
POPULATION SURVEY AMÉRICAIN

6.1 Plan d'échantillonnage de la CPS

La CPS est une enquête-ménage mensuelle entreprise par le Census Bureau des Etats-Unis en collaboration avec le Bureau of Labor Statistics afin de fournir une estimation nationale des caractéristiques de la population active, tels le nombre de travailleurs, de chômeurs et de personnes actives, ainsi que d'autres particularités de la population civile non institutionnalisée. Le plan d'échantillonnage de la CPS prévoit le renouvellement d'une fraction des ménages de l'échantillon tous les mois. Chaque mois, l'échantillon se compose de huit panels ou groupes de renouvellement, l'un d'eux faisant l'objet d'une première interview, le second d'une deuxième interview, ... et le dernier d'une huitième interview. En d'autres mots, le programme d'interviews s'articule sur le temps passé par le panel dans l'échantillon. Les ménages d'un groupe sont interviewés pendant quatre mois consécutifs, oubliés pendant huit mois puis interviewés de nouveau quatre mois de suite. Ils sont alors rayés de l'échantillon. C'est ce qu'on appelle un plan de renouvellement 4-8-4. Il s'agit d'un cas particulier des plans décrits par Rao et Grahm (1964).

6.2 Estimation et méthodes d'estimation de la variance

Nous nous sommes servis des estimations de la covariance des données venant de la CPS pour comparer les différents estimateurs et plans de renouvellement. Adam et Fuller (1992) ainsi que Fuller, Adam et Yansaneh (1993) décrivent en détail la construction du modèle, l'estimation de ses paramètres et l'estimation de la covariance des observations pour un groupe de renouvellement donné à l'égard de diverses caractéristiques intéressantes. Puisqu'ils viennent du même jeu d'unités primaires d'échantillonnage, les groupes de renouvellement ne sont pas indépendants et la covariance intègre un élément pour refléter le fait que les unités primaires d'échantillonnage ne changent pas. On calcule l'EBR au moyen des huit estimateurs simples courants et des 15 estimateurs des 15 périodes antérieures. Pour cela, on se sert de la covariance afin de créer huit combinaisons linéaires des observations courantes et des quinze qui les précèdent, sans corrélation avec les quinze observations antérieures. À cause de l'effet des unités primaires d'échantillonnage, les combinaisons linéaires sont corrélées avec les observations de la même bande situées à plus de 15 périodes dans le passé. Elles présentent donc une corrélation avec les estimateurs antérieurs. La matrice de covariance inclut les corrélations avec les estimateurs antérieurs $\theta_{t-1}, \dots, \theta_{t-15}$, $t = 1, \dots, 15$ à la construction de l'estimateur de θ_t . Néanmoins, puisqu'on ne retient que les 15 observations les plus récentes, l'estimateur de θ_t résultant ne donne pas le MELSB du niveau courant. La matrice de covariance de $(\theta_{t-15}, \dots, \theta_{t-1}, \theta_t)'$ est exacte et on estime que

6.3 Résultats numériques et analyse
6.3.1 Comparaison des estimateurs

L'étude à celle de l'estimateur classique du niveau courant pour chaque caractéristique digne d'intérêt. On se appellera que l'estimateur de base du niveau courant, noté \bar{y}_t , correspond à la simple moyenne des huit estimateurs élémentaires venant des huit groupes de renouvellement observés au temps t , c'est-à-dire $\bar{y}_t = 8^{-1} \sum_{k=1}^8 y_{t,k}$, et $\text{Var}(\bar{y}_t) = \sigma^2/8$, où $\sigma^2 = \text{Var}(y_{t,k})$ pour toutes les valeurs t et k . L'estimateur de base du changement survenu entre deux périodes correspond à l'écart entre la simple moyenne des deux périodes concernées.

Le tableau 1 compare la variance des estimateurs à celle de l'estimateur classique du niveau courant pour chaque caractéristique digne d'intérêt. On se appellera que l'estimateur de base du niveau courant, noté \bar{y}_t , correspond à la simple moyenne des huit estimateurs élémentaires venant des huit groupes de renouvellement observés au temps t , c'est-à-dire $\bar{y}_t = 8^{-1} \sum_{k=1}^8 y_{t,k}$, et $\text{Var}(\bar{y}_t) = \sigma^2/8$, où $\sigma^2 = \text{Var}(y_{t,k})$ pour toutes les valeurs t et k . L'estimateur de base du changement survenu entre deux périodes correspond à l'écart entre la simple moyenne des deux périodes concernées.

Nous nous attarderons seulement à l'estimation des paramètres de deux caractéristiques: les personnes occupées et les chômeurs. Les paramètres auxquels nous nous intéressons dans chaque cas sont le niveau courant et le changement d'une période à l'autre, pendant un maximum de 12 périodes. Les estimateurs faisant l'objet de la comparaison sont l'estimateur composite de la CPS, l'EBR et le MELSB obtenu avec 2, 3, 12, 16 et 24 périodes, la MELSB de 12 et de 16 périodes. Comme le fait le Bureau of Labor Statistics américain avec les estimateurs de la CPS, les estimateurs ne sont pas modifiés en fonction des nouvelles données. L'estimateur du changement de niveau d'une caractéristique à laquelle on s'intéresse entre les temps $t-1$ et t ne constitue donc pas l'estimateur idéal à l'ensemble des données disponibles. L'estimateur idéal serait l'écart entre le meilleur estimateur au temps t d'après les données recueillies jusqu'à ce moment, et le meilleur estimateur au temps $t-1$, selon les données recueillies jusqu'au temps $t-1$.

Nous ne tiendrons pas compte de la désaisonnalisation dans l'analyse. Les méthodes d'estimation présentées ici peuvent toutefois s'y appliquer. Pour établir la variance d'un estimateur donné à un moment précis dans le temps, on exprime d'abord l'estimateur sous forme de combinaison linéaire des observations existantes au moment concerné. On calcule ensuite la variance de l'estimateur en fonction des coefficients de la combinaison linéaire et des éléments de la matrice de covariance.

Le MELSB reposant sur 3 périodes ou plus donne de meilleurs estimateurs du niveau courant que l'estimateur composite de la CPS. En général, la meilleure méthode d'estimation linéaire sans biais voit son efficacité statistique augmenter avec le nombre de périodes. En ce qui concerne nos deux caractéristiques, les résultats montrent que la meilleure méthode d'estimation linéaire sans biais reposant sur 12 périodes donne toujours des résultats supérieurs à

indépendantes sur les s bandes à l'instant présent. Soit les observations transformées, notées $z_{i,c}^c, i = 1, 2, \dots, s$,

$$z_{i,c}^c = y_{i,c}^c - \sum_{j=1}^J b_{k(i,c),j} y_{i,c-j}^c \quad (6)$$

où $b_{k(i,c),j}$ sont des coefficients de telle sorte que $z_{i,c}^c$ ne présente aucune corrélation avec $y_{i,c-j}^c$ pour toutes les valeurs $j > 0$. En raison des hypothèses (4) et (5), les coefficients $b_{k(i,c),j}$ sont fixes dans le temps. De par l'hypothèse (3), $z_{i,c}^c$ ne présente pas de corrélation avec les observations antérieures. La valeur probable de $z_{i,c}^c$ est

$$\theta_{i,c}^c - \sum_{j=1}^J b_{k(i,c),j} \theta_{i,c-j}^c, i = 1, 2, \dots, s.$$

5.2 L'estimateur de régression récursif

Soit $\theta_h(t), h \leq t$, désigne l'estimateur des moindres carrés du paramètre (scalaire) θ_h construit grâce aux données recueillies jusqu'au temps t et soit $\Theta_{i(m)}^{t-1} = (\theta_{i(m)}^{t-1}(t), \dots, \theta_1^{t-1}(t))'$, l'estimateur des moindres carrés du vecteur des m paramètres $\theta_{i(m)}^{t-1}, \dots, \theta_1^{t-1}$, au temps t , obtenu grâce aux données recueillies jusqu'au temps t . L'objectif consiste à bâtir l'estimateur à variance minimale de $\theta_{i,c}^c$, le niveau courant du paramètre auquel on s'intéresse, au moyen de toutes les données disponibles au temps c . Voici un modèle linéaire des données existantes à l'instant présent

$$Z^c = W \Theta_{i(m)}^{c(m+1)} + U^c \quad (7)$$

où

$$W = \begin{pmatrix} I_m & X_{21} \\ 0 & J_s \end{pmatrix}$$

$Z^c = (\theta_{i(m)}^{c-1}(m), z_{i,c}^c, z_{i,c}^c, \dots, z_{i,c}^c)$, et X_{21} représente une $s \times m$ matrice aux éléments constants dans le temps, fonction des coefficients $b_{k,j}$ de (6). Si $\text{Var}\{z_{i,c}^c\} = \sigma_{i,c}^2, i = 1, 2, \dots, s$, et Ω_{22} est la matrice diagonale composée des éléments diagonaux $\sigma_{i,c}^2$, la matrice de covariance de Z^c est $V^c = \text{blockdiag}\{\sum_{c=1}^{c-1} \Omega_{22}, \Omega_{22}\}$. On présume que $\sigma_{i,c}^2, i = 1, 2, \dots, s$, sont des valeurs positives.

L'estimateur de régression récursif (ERR) de $\Theta_{i(m)}^{c(m+1)}$ correspond à l'estimateur des moindres carrés de $\Theta_{i(m)}^{c(m+1)}$ selon le modèle (7). Par conséquent, l'ERR de $\Theta_{i(m)}^{c(m+1)}$ est

$$\Theta_{i(m)}^{c(m+1)} = (W' V^{-1} V^{-1} W')^{-1} W' V^{-1} Z^c \quad (8)$$

L'estimateur (8) trouve son utilité dans la simplicité des calculs. À n'importe quel moment t de l'enquête répétitive, on peut extraire l'information nécessaire à l'estimation de $\theta_1, \theta_{i-1}, \dots, \theta_{i-m}$ d'un ensemble de m estimations récursives des moindres carrés et des observations courantes.

Nous approfondirons maintenant la méthode de régression récursive. Au temps t , on a $\Theta_{i(m)}^{t(m+1)}$, l'ERR de $\Theta_{i(m)}^{t(m+1)}$ au temps t et sa $(m+1) \times (m+1)$ matrice de covariance $\sum_{i(m+1)}^{t(m+1)}$ en

$$\sum_{i(m+1)}^{t(m+1)} = \begin{pmatrix} v_{11,t} & v_{12,t} \\ v_{12,t} & v_{22,t} \end{pmatrix} = \begin{pmatrix} \sum_{i(m)}^{t-1} V_{11,i} & \sum_{i(m)}^{t-1} V_{12,i} \\ \sum_{i(m)}^{t-1} V_{12,i} & \sum_{i(m)}^{t-1} V_{22,i} \end{pmatrix}$$

où $v_{11,t}$ représente la variance de $\theta_{i(m)}^{t-1}(t)$, $\sum_{i(m)}^{t-1} \theta_{i(m)}^{t-1}(t)$, et $V_{12,i}$ correspond à la covariance entre ces deux quantités. Signalons que si on garde $\theta_{i(m)}^{t-1}$ dans le vecteur paramétrique et $\theta_{i(m)}^{t-1}$ dans le vecteur des données, l'estimateur de $\theta_{i(m)}^{t-1}$ ne change pas (normalement, il le ferait). On le doit au fait que l'estimateur du vecteur original d'un problème où interviennent les moindres carrés ne varie pas quand s'y ajoute une observation dont la probabilité est égale à celle d'un seul nouveau paramètre. Pour actualiser l'ERR la période suivante, on retranche donc l'estimation initiale de la plus ancienne période $\theta_{i(m)}^{t-1}(t)$ du vecteur des données et le paramètre correspondant $\theta_{i(m)}^{t-1}$ du vecteur paramétrique. On ajoute ensuite à ce dernier le paramètre $\theta_{i(m)}^{t-1}$. De cette manière, la matrice W du modèle de base dans le problème d'estimation garde le même nombre de dimensions dans le temps. Pour la catégorie d'enquêtes répétitives examinée ici, les calculs requis pour obtenir le MELSB du vecteur des paramètres auxquels on s'intéresse ne dépasseront donc pas une certaine limite.

Au temps $t+1$, le modèle prend la forme du modèle (7) où $c = t+1$, $Z^{t+1} = (\theta_{i(m)}^{t-m+1}(t), \theta_{i-1}^{t-m+1}(t), z_{i,t+1}^{t+1}, \dots, z_{i,t+1}^{t+1})'$, et la matrice de covariance de Z^{t+1} est $V^{t+1} = \text{blockdiag}\{\sum_{i(m)}^{t-1} \Omega_{22}, \Omega_{22}\}$. Le MELSB de $\Theta_{i(m)}^{t+1(m+1)}$ et sa matrice de covariance dérivent des formules habituelles des moindres carrés. On se sert ensuite des estimateurs des moindres carrés des m derniers éléments de $\Theta_{i(m)}^{t+1(m+1)}$ comme estimations initiales de l'itération suivante dans le modèle.

La matrice de covariance du vecteur des estimateurs récursifs des moindres carrés converge vers une matrice positive définie à mesure que le nombre de périodes de l'enquête tend vers l'infini. On en présente la preuve à l'annexe.

Théorème: À n'importe quel moment t , supposons que le vecteur des estimateurs récursifs des moindres carrés $\Theta_{i(m)}^{t(m+1)} = (\theta_{i(m)}^{t-m+1}(t), \theta_{i-1}^{t-m+1}(t), \theta_1^{t(m)}(t))'$ corresponde au MELSB du vecteur des paramètres $\Theta_{i(m)}^{t(m+1)} = (\theta_{i(m)}^{t-m+1}, \theta_{i-1}^{t-m+1}, \dots, \theta_1^{t(m)})'$ reposant sur les données recueillies jusqu'au temps t . Soit $\sum_{i(m)}^{t(m+1)}$, la matrice de covariance de $\Theta_{i(m)}^{t(m+1)}$. Supposons que les hypothèses (1) à (5) se vérifient. Supposons aussi que les éléments de $V^{t(m+1)}$ sont bornés pour toutes les valeurs n , où V^n représente la matrice de covariance de n observations. Il s'ensuit que la matrice de covariance $\sum_{i(m)}^{t(m+1)}$ converge lorsque $t \rightarrow \infty$; et la limite correspond à une $m \times m$ matrice positive définie.

amorçé au temps zéro

(3)

période courante.

SANS BIAIS

périodes.

partir des données existantes à cet instant.

modèle linéaire de Y^d est

(5)

Markov, le MELSB de Θ^d est

à correspond

$${}^dA_{\Gamma} {}^dA_{\Gamma} X_{\Gamma} (X_{\Gamma} {}^dA_{\Gamma} X) = {}^d\bigoplus_{\Gamma}$$

RECURSIVE

On s'est rendu compte que les techniques d'estimation récurrentes ont leur utilité quand les données ne sont pas toutes disponibles en même temps, mais s'accrément à la longue, et quand il s'avère peu pratique de calculer une estimation optimale à partir de l'ensemble des données. On lira Odell et Lewis (1971), Sallais et Harville (1981) ainsi que les auteurs cités en référence par ces chercheurs, pour connaître les algorithmes récurrents de la meilleure estimation linéaire sans biais. Tiller (1989) propose pour sa part l'approche du filtre de Kalman à l'estimation des caractéristiques de la population active à partir des données d'enquête. Ainsi qu'on a pu le voir à la partie 4, le calcul direct du MBLSB se complique peu à peu à mesure que le nombre de périodes augmente. Nous proposons une méthode d'estimation par régression récursive faisant appel à un jeu d'estimations initiales judicieusement choisies, aux nouvelles observations du niveau courant et aux observations antérieures sur les groupes de renouvellement.

et proposition d'un estimateur

Supposons qu'une enquête fonctionne depuis au moins

(3) les groupes de renouvellement sont indépendants;

tante dans le temps pour la même bande, et est la même

Ces hypothèses nous serviront à bâtir un estimateur linéaire. On assurplira l'hypothèse (3) pour calculer la variance de l'estimateur. En vertu des hypothèses (1) et (3), les observations séparées de plus de m périodes sont indépendantes. À l'instant présent, noté c , où $c > m$, on observe un jeu s d'estimateurs élémentaires du paramètre θ^c . Pour obtenir l'estimateur généralisé des moindres carrés, on transforme les s observations courantes afin qu'elles ne présentent pas de corrélation avec les observations antérieures. Après transformation, les valeurs probables des observations constituent une fonction de θ^c et des paramètres des m périodes antérieures. Supposons qu'on connaisse le MELSB du vecteur de paramètres des m périodes antérieures et la matrice de covariance $m \times m$ de ces estimateurs. Au temps c , on a donc: i) m estimations initiales $\Theta^{c-1(m)} = (\theta^{c-m}, \dots, \theta^{c-1})'$; et ii) la matrice de covariance $\sum \Sigma^{c-1(m)}$ de $\Theta^{c-1(m)}$ et iii) s observations

où, pour l'estimateur courant, $\pi = 0,4$ et $\pi_2 = 0,2$, $y_{t,k}$ correspond à l'estimation du niveau venant du groupe de renouvellement qui se retrouve pour la k -ième fois dans l'échantillon au temps t , $\bar{y}_t = 8^{-1} \sum_{k=1}^8 y_{t,k}$ représente l'estimateur de base, soit la moyenne des estimations élémentaires issues des huit groupes de renouvellement au temps t , $\theta_{t-1,c}$ est l'estimateur composite au temps

$$\theta_{t,c} = (1 - \pi_1) \bar{y}_t + \pi_1 (\theta_{t-1,c} + \delta_{t-1,c} + \pi_2 \delta_{t-1,c}) \quad (1)$$

déterminé par deux paramètres et s'écrit
L'estimateur composite actuellement en usage est baptisé «nouveaux» groupes de renouvellement.

En général, les estimateurs composites combinent un ou plusieurs estimateurs récents et les données des périodes courante et antérieure(s) pour donner l'estimateur de la période courante. Dans la CPS, six des huit groupes de renouvellement observés au temps t ont été également au temps $t-1$. Nous les appellerons les groupes de renouvellement permanents, tandis que les deux autres seront baptisés «nouveaux» groupes de renouvellement.

3. L'ESTIMATEUR COMPOSITE DE LA CPS

Les enquêtes comme la CPS et l'Enquête sur la population active canadienne satisfont ces hypothèses. Yansaneh (1997) donne une illustration du plan de renouvellement 4-8-4 utilisé par la CPS.

première par m périodes.
dernière, l'ultime observation étant séparée de la première fois, un autre pour la deuxième, ..., un pour la m à un moment quelconque, un est observé pour la s groupes de renouvellement que comprend l'échantillon passé dans l'échantillon. En d'autres termes, parmi les s groupes de renouvellement est équilibré avec le temps

2) Le plan d'échantillonnage est équilibré avec le temps d'un groupe de renouvellement à l'autre.
 $m+1$; les observations suivent un ordre fixe, identique, une bande pendant une période dont la durée totale égale

1) On observe un groupe de renouvellement donné dans une bande pendant une période dont la durée totale égale $m+1$; les observations suivent un ordre fixe, identique, retrouver dans une bande. Posons les hypothèses que voici:
«bandes». Plusieurs groupes de renouvellement peuvent se retrouver dans une bande. Posons les hypothèses que voici:
groupe de renouvellement ne se retrouvent que dans une colonne. Le nombre total d'estimateurs élémentaires sera $n = p \times s$. Les colonnes de matrices H sont appelées «bandes». Plusieurs groupes de renouvellement peuvent se retrouver dans une bande. Posons les hypothèses que voici:
groupe de renouvellement ne se retrouvent que dans une colonne. Le nombre total d'estimateurs élémentaires sera $n = p \times s$. Les colonnes de matrices H sont appelées «bandes». Plusieurs groupes de renouvellement peuvent se retrouver dans une bande. Posons les hypothèses que voici:

et

$$\delta_{t,c} = 8^{-1} \left(\sum_{k \in T} y_{t,k} - 3^{-1} \sum_{k \in S} y_{t,k} \right),$$

$$\delta_{t-1,c} = 6^{-1} \sum_{k \in S} (y_{t,k} - y_{t-1,k-1}),$$

$t-1$, $\delta_{t-1,c}$ estime le changement de niveau d'après les six groupes permanents au temps t , et $\delta_{t,c}$ donne l'écart entre la moyenne des deux nouveaux groupes de renouvellement et celle des six permanents. Par conséquent,

$$\theta_{t,c} = \sum_{k=1}^8 \omega_{1,k(t)} y_{t,k} + \sum_{k=9}^{16} \omega_{2,k(t)} y_{t-1,k-1} + \pi_1 \theta_{t-1,c} \quad (2)$$

où $k(t)$ est le temps passé par l'observation (ii) dans l'échantillon, en fonction de la bande (t) et du temps (t). Si $\lambda_1 = 1/8$ et $\lambda_2 = -1/6$, et $\lambda_3 = 1/3$, alors $\omega_{2,k} = \pi_1 \lambda_2$, et $\omega_{1,k} = (1 - \pi_1) \lambda_2 - \pi_1 \lambda_3$ pour $k \in S$ et $\omega_{1,k} = \lambda_1 (1 - \pi_1 + \pi_2)$ pour $k \in T$. Soit $y_{t,i}$, $i = 1, 2, \dots, s$, l'estimateur élémentaire du paramètre auquel on s'intéresse, tiré du groupe de renouvellement se trouvant dans la bande i au temps t . L'estimateur composite de la CPS prend la forme

$$\theta_{t,c} = \sum_{k=1}^8 \omega_{1,k(t)} y_{t,k} + \sum_{k=9}^{16} \omega_{2,k(t)} y_{t-1,k-1} + \pi_1 \theta_{t-1,c}$$

Soit

$$D_1 = (\omega_{1,k(1)}, \omega_{1,k(2)}, \dots, \omega_{1,k(8)})'$$

$$D_2 = (\omega_{2,k(1)}, \omega_{2,k(2)}, \dots, \omega_{2,k(8)})'$$

et $y' = (y_{1,1}, y_{1,2}, \dots, y_{8,1})'$. Dans ce cas,

Méthode optimale d'estimation réursive pour les enquêtes répétitives

IBRAHIM S. VANSANEH et WAYNE A. FULLER¹

RÉSUMÉ

Les auteurs évaluent la méthode d'estimation par les moindres carrés pour les enquêtes répétitives. Ils proposent plusieurs estimateurs pour le niveau courant, le changement de niveau et le niveau moyen applicables à des périodes multiples. Suit la présentation de l'estimateur de régression récurrent, méthode réursive permettant de calculer le meilleur estimateur linéaire sans biais d'après l'ensemble des périodes couvertes par l'enquête. On constate qu'il y a convergence de la régression réursive et que le nombre de dimensions de l'estimation est plafonné lorsque le nombre de périodes tend vers l'infini. La méthode réursive apporte une solution au problème de la complexité des calculs que suscite l'estimation non biaisée de la variance minimale dans les enquêtes répétitives. Les auteurs recourent aux données de la U.S. Current Population Survey pour comparer les différents estimateurs, avec deux genres de plan d'échantillonnage: le plan à renouvellement intermittent de la Current Population Survey et deux plans à renouvellement continu.

MOTS CLÉS: Estimation de régression réursive; estimation composite; plans d'échantillonnage avec renouvellement; groupes de renouvellement.

1. INTRODUCTION

Nous nous pencherons sur la méthode d'estimation par les moindres carrés utilisée de façon répétitive, avec chevauchement partiel des unités d'échantillonnage. Duncan et Kalton (1987) traitent en général des différents types d'enquête et des objectifs de ces dernières. Nous nous intéresserons ici aux enquêtes dont le panel est renouvelé, et où on procède à des déterminations répétitives sur certaines unités d'échantillonnage, sans que chacune revienne pour autant constamment dans l'échantillon.

C'est Patterson (1950) qui, s'inspirant des travaux de Cochran (1942) et de Jessen (1942), a posé les fondements théoriques du plan d'échantillonnage et de la méthode d'estimation des enquêtes répétitives articulée sur la méthode généralisée des moindres carrés. Plusieurs autres auteurs ont approfondi cette dernière. On lira notamment Fuller (1990) et d'autres, qu'il cite en référence. Vansaneh (1992) s'est intéressé à une méthode d'estimation par les moindres carrés applicable à une catégorie assez générale d'enquêtes répétitives. L'estimation composite est une méthode d'estimation destinée aux enquêtes répétitives en vertu de laquelle on se sert des observations de la période en cours et de celle qui la précède, ainsi que de l'estimateur de niveau de la période antérieure. Breaun et Ernst (1983) ont comparé quelques estimateurs à l'estimateur composite de la U.S. Current Population Survey (CPS). Kumar et Lee (1983) en ont fait autant avec les données de l'Enquête sur la population active (EPA) canadienne. De son côté, Wolter (1979) a proposé une stratégie complète pour l'estimation composite générale des plans de renouvellement à deux degrés comme celui de la Retail Trade Survey du Census Bureau des États-Unis. Enfin, Singh (1996) a suggéré l'usage d'une autre catégorie d'estima-

Nous aborderons ici la question des méthodes d'estimation des enquêtes répétitives selon l'hypothèse que les valeurs réelles inconnues sont des paramètres fixes. Les estimateurs sont comparés à la méthode d'estimation composite dont on se sert présentement dans la CPS. L'article se présente comme suit. La partie 2 formule quelques hypothèses de base sur la catégorie générale d'enquêtes répétitives examinée. À la partie 3, on trouvera une description de la méthode d'estimation composite de la CPS. Suit une analyse de la méthode de la meilleure estimation linéaire sans biais, à la partie 4. Nous présentons une méthode d'estimation par régression réursive à la partie 5. Cette méthode a pour but de simplifier les calculs associés à la meilleure estimation linéaire non biaisée. Enfin, la partie 6 applique la méthode aux données de la CPS. On y compare d'autres estimateurs et plans d'échantillonnage avec renouvellement.

2. HYPOTHÈSES FONDAMENTALES

Cette partie décrit les enquêtes du genre qui nous intéresse. Un groupe de renouvellement est un groupe de personnes choisies pour constituer un échantillon et observées pendant un nombre déterminé de périodes, selon un schéma fixe dans le temps. Supposons qu'au cours de

¹ Ibrahim S. Vansaneh, Statistical Group, Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850; et Wayne A. Fuller, Department of Statistics, Iowa State University, Ames, IA 50011 U.S.A.

Les auteurs aimeraient remercier le personnel de la Cinergy Corporation pour leur avoir donné l'occasion de mener l'enquête sur les caractéristiques des immeubles, qui est à l'origine de la présente recherche. Nous voulons également remercier les trois examinateurs pour leurs excellentes suggestions qui ont contribué à améliorer sensiblement ce document.

REMERCIEMENTS

BIBLIOGRAPHIE

- BANDYOPADHYAY, S., et ADHIKARI, A.K. (1993). Échantillon-nage dans des bases imparfaites contenant un nombre inconnu d'enregistrements répétés. *Techniques d'enquête*, 19, 205-209.
- BIRNBAUM, Z.W., et SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, PHS Publication 1000, Ser. 2, *Data Evaluation and Methods Research*, no. 11. Hyattsville, MD: National Center for Health Statistics, Public Health Service.
- U.S. Department of Health and Human Services.
- CASADY, R.J., et SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-605.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3ième éd.). New York: Wiley & Sons.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953a). *Sample Survey Methods and Theory I, Methods and Applications*. New York: Wiley & Sons.
- HANSEN, M.H., HURWITZ, W.N., et MADOW, W.G. (1953b). *Sample Survey Methods and Theory 2, Theory*. New York: Wiley & Sons.
- HORVITZ, D.G., et THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.
- KISH, L. (1965). *Survey Sampling*. New York: Wiley & Sons.
- LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.
- LESSLER, J.T., et KALSBEEK, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley & Sons.
- MUSSER, O. (1993). Unbiased estimation in the presence of frame duplication. *Proceedings of the International Conference on Establishment Surveys*, 889-892.
- SIRKEN, M.G. (1972a). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.
- SIRKEN, M.G. (1972b). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 65, 224-227.
- U.S. DEPARTMENT OF ENERGY, Energy Information Administration (1992). *Commercial Buildings Energy Consumption Survey*.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.
- WRIGHT, T., et TSAO, H.J. (1983). A frame on frames: An annotated bibliography, (Ed., Tommy Wright). *Statistical Methods and the Improvement of Data Quality*, Orlando, Florida: Academic Press, 25-72.

figuraient dans la base de sondage et si elles étaient ou non reliées à d'autres éléments de la population (immeubles commerciaux). Dans le cas de scénarios plus complexes, les intervieweurs ont parfois eu recours à des esquisses des immeubles et à l'étiquetage de toutes les adresses pertinentes, ce qui nous a permis de déterminer la structure de tous les sous-graphes CM dans notre échantillon et de déterminer les facteurs de pondération appropriés (s_k^*).

Nous avons également défini les formules pour calculer la variance de certains des estimateurs présentés dans le document. Il convient de préciser que ces formules sont des paramètres de la population qui ne peuvent être converties facilement en estimations de l'échantillon correspondant. En fait, les auteurs ne connaissent aucune méthode optimale pour estimer les variances décrites dans cet article. Il existe cependant de nombreuses méthodes exigeant beaucoup de calculs (méthode BRR, méthode «bootstrap», etc.) pour estimer la variance dans le cadre de sondages complexes (Wolter 1985). Il y a lieu également de préciser que chacune de ces méthodes d'estimation de la variance vise un objectif commun, à savoir les formules de la variance que nous avons élaborées ici.

L'utilité de ces formules réside toutefois dans leur application à l'étude des effets des imperfections de la base de sondage, et des caractéristiques de la population, sur la précision des estimations. Une telle étude, qui pourrait faire l'objet d'une autre recherche, devrait mener à la formulation de recommandations et de lignes directrices à l'intention des chercheurs, sur la façon de gérer une base de sondage de structure plusieurs à plusieurs. En d'autres mots, le chercheur devrait, à partir de la base de sondage et des caractéristiques de la population, être en mesure de prendre des décisions stratégiques sur les options qui s'offrent: recenser la population par interview pour supprimer les imperfections de correspondance ou utiliser les estimateurs décrits dans le présent article.

Un autre domaine de recherche futur serait de comparer la précision de nos estimateurs à celle d'autres estimateurs, par exemple l'estimateur de Horvitz-Thompson. Comme nous l'avons indiqué dans l'introduction, l'estimateur de Horvitz-Thompson peut être appliqué à une méthode d'échantillonnage basée sur une structure de plusieurs à plusieurs. L'avantage de l'estimateur de Horvitz-Thompson est qu'il permet, moyennant des probabilités d'inclusion de premier et de second ordres bien définies, d'obtenir à la fois une estimation des caractéristiques de la population et une estimation sans biais de sa variance. En outre, les probabilités d'inclusion du premier ordre peuvent être calculées d'une manière similaire à celle de Musser (1993) à partir uniquement d'information provenant des sous-graphes CM. Ces probabilités sont toutefois très difficiles à calculer dans une base de sondage complexe de structure plusieurs à plusieurs comme la nôtre. Il est en revanche relativement facile de calculer les facteurs de pondération nécessaires pour nos estimateurs.

$$S_2^h = E(Y_2^h)^2 = E(Y_2^{*h})^2 - (Y_2^{*h})^2 + V(Y_2^{*h}) = E\left(\sum_{l=1}^L Y_2^h\right)^2 - \sum_{l=1}^L Y_2^h$$

$$\sum_{l=1}^L (Y_2^h)^2 = \sum_{l=1}^L E(Y_2^h)^2 - (Y_2^h)^2 + \sum_{l=1}^L S_2^h.$$

Maintenant,

$$E(Y_2^h)^2 = \frac{N_2^h}{2} E\left(\sum_{i=1}^n x_{hK_i}\right)^2 =$$

$$\frac{N_2^h}{2} \left(\sum_{i=1}^n E(x_{hK_i})^2 + 2E\left(\sum_{i < j} x_{hK_i} x_{hK_j}\right) \right). \quad (3.7)$$

Pour chaque $i = 1, \dots, n_h$,

$$E(x_{hK_i})^2 =$$

$$\sum_{k \in V_h} \left(\left(\frac{y_k}{s_k} \right)^2 \Pr(hK_i = k) \right) = \sum_{k \in V_h} \left(\left(\frac{y_k}{s_k} \right)^2 \frac{q_{hk}}{N_h} \right). \quad (3.8)$$

En utilisant les équations (2.7) et (2.8),

$$2E\left(\sum_{i > i'} x_{hK_i} x_{hK_{i'}}\right) = 2 \left(\frac{2}{n_h} \right) E(x_{hK_i} x_{hK_{i'}}) = \sum_{[hjK, hj'K'] \in R_h^*} \frac{y_k}{s_k} \frac{y_{k'}}{s_{k'}} \Pr(hK_i = k, hK_{i'} = k') =$$

$$n_h(n_h - 1) \sum_{[hjK, hj'K'] \in R_h^*} \left[\left(\frac{y_k}{s_k} \frac{y_{k'}}{s_{k'}} \right) \left(\frac{2}{N_h} \right)^{-1} \frac{1}{s_{jk's'k'}} \right]$$

$$2n_h(n_h - 1) \sum_{[hjK, hj'K'] \in R_h^*} \left[\frac{y_k s_{hjK}}{s_k} \frac{y_{k'} s_{hj'K'}}{s_{k'}} \right]. \quad (3.9)$$

L'équation (3.5) découle maintenant de (3.8), (3.9) et de la définition de Y_h^* .
En utilisant la méthode du corollaire 2-1, l'équation (3.5) peut être simplifiée aux fins de calcul, comme suit:

4. CONCLUSIONS

Comme dans le cas de l'EASSR, l'estimateur de la moyenne d'une population est biaisé, parce qu'il s'agit d'un estimateur par quotient.

$$\hat{Y}_{st} = \sum_{h=1}^L \frac{N_h}{N} \hat{Y}_h, \text{ où } \hat{Y}_h = \frac{\sum_{i=1}^n x_{hK_i}}{\sum_{i=1}^n \frac{1}{s_{hK_i}}}. \quad (3.10)$$

L'estimateur de la moyenne d'une population, selon un plan d'échantillonnage aléatoire stratifié avec base de sondage de plusieurs à plusieurs, est représenté par:

L'estimateur élaboré ici pour la moyenne d'une population selon l'échantillonnage aléatoire stratifié constitue un prolongement de l'estimateur proposé par Hansen et coll. 1953a (p. 62-64), ici appliqué à un échantillon aléatoire stratifié prélevé d'une base de plusieurs à plusieurs.

3.4 Moyennes de la population

3.4.1 Estimateur de la moyenne d'une population

où A_h représente l'ensemble des arcs qui partent des unités dans F_h .

$$S_2^h = \frac{N_h}{2} \left[\sum_{k \in V_h} q_{hk} \left(\frac{y_k}{s_k} \right)^2 + \frac{(N_h - 1)}{2} \left(\sum_{[hjK \in A_h]} \frac{y_k s_{hjK}}{s_k} \frac{y_{k'} s_{hj'K'}}{s_{k'}} \right)^2 - \sum_{[hjK, hj'K'] \in R_h^*} \left(\frac{y_k s_{hjK}}{s_k} \frac{y_{k'} s_{hj'K'}}{s_{k'}} \right)^2 \right] -$$

La méthode de pondération décrite ici a été appliquée à une étude sur les immeubles commerciaux pour laquelle un échantillon aléatoire stratifié avait été utilisé. Pour cette étude, pour laquelle la base de sondage était formée des adresses de voirie, les intervieweurs ont noté toute adresse additionnelle se rapportant à l'immeuble sélectionné. Il a ensuite été déterminé si ces adresses additionnelles

L'estimateur du total d'une population, selon un plan d'échantillonnage aléatoire stratifié avec une base de sondage de structure plusieurs à plusieurs, est:

$$\hat{Y}^{st} = \sum_{h=1}^H \hat{Y}_h, \text{ où } \hat{Y}_h = \frac{n_h}{N_h} \sum_{i=1}^{n_h} x_{hk_i}. \quad (3.1)$$

3.3.2 Variance de l'estimateur pour le total d'une population

Avant de définir la variance de l'estimateur (3.1),

quelques termes additionnels doivent être définis. Supposons que q_{hk} représente le «facteur de pondération de l'élément de la strate». Ce facteur additionnel est nécessaire en raison du risque de chevauchement. Supposons maintenant que U_{hk} représente l'ensemble des unités dans F_h dont les arcs se terminent à l'élément de population k , par exemple $U_{24} = \{(2, 1), (2, 2)\}$. Définissons maintenant

$q_{hk} = \sum_{h_j \in U_{hk}} s_{hj_k}$. Pour illustrer ceci, rappelons-nous, à la figure 3.1, que l'élément de la population 4 est représenté par deux unités de la base dans la strate 2, de sorte que

$$q_{24} = \sum_{2j \in U_{24}} s_{2j4} = 2.$$

Le facteur de pondération q_{hk} joue le rôle de s_k lorsque la sélection se limite à F_h . En fait, $q_{hk} = s_k$ lorsque'il n'y a pas de chevauchement. La probabilité de sélectionner quelque unité de F_h à l'étape i sur n_h est égale à $1/N_h$. Cependant, la probabilité de sélectionner un élément de la population k représenté par une unité dans F_h est égale à

$$\Pr(hK_i = k) = q_{hk}/N_h, \text{ pour toutes les } i = 1, \dots, n_h.$$

Pour faire la preuve, nous introduisons le terme «total réparti entre les strates», représenté par Y_h^* . En fait, les valeurs des éléments de la population qui sont représentés par les unités dans les strates multiples sont réparties entre ces strates. Supposons que V_h représente l'ensemble des éléments de la population associés aux unités dans F_h . Dans notre exemple, $V_1 = \{1, 2, 3, 4\}$ et $V_2 = \{4, 5, 6, 7\}$. Supposons que

$$Y_h^* = \sum_{k \in V_h} Y_k q_{hk} / s_k,$$

où Y_k est la valeur de l'élément de population k , $k = 1, 2, \dots, M$. Lorsqu'il y a chevauchement, l'utilisation des facteurs de pondération q_{hk} et s_k réparti la mesure Y_k entre les strates dans lesquelles l'élément de population k est représenté. L'utilisation de ces facteurs de pondération sert en fait à répartir la valeur de l'élément de population entre les strates, en fonction du nombre de fois que cet élément est représenté dans une strate par rapport au nombre total de fois qu'il est représenté dans la base de sondage. À la figure 3.1, par exemple, Y_1^* et Y_2^* se calculent comme suit:

$$Y_1^* = \frac{30(1)}{15(1/2)} + \frac{1}{5(1/2)} + \frac{4}{65(2)} = 82,5$$

$$Y_2^* = \frac{4}{65(2)} + \frac{1}{10(3/2)} + \frac{1/2}{5(1/2)} + \frac{1/2}{20(2)} = 67,5.$$

À noter que $\sum_{h=1}^H Y_h^* = Y$, qu'il y ait ou non chevauchement. **Théorème 3-1:** L'estimateur du total d'une population (3.1) est sans biais.

Preuve:

À partir de (3.1),

$$E(\hat{Y}^{st}) = \sum_{h=1}^H \sum_{i=1}^{n_h} \frac{n_h}{N_h} E(x_{hk_i}). \quad (3.2)$$

Pour chaque $i = 1, \dots, n_h$,

$$E(x_{hk_i}) = \sum_{k \in V_h} \frac{Y_k}{N_h} \Pr(hK_i = k) =$$

$$\sum_{k \in V_h} \frac{Y_k}{N_h} \frac{q_{hk}}{N_h} = \frac{1}{N_h} \sum_{k \in V_h} Y_k q_{hk} = \frac{1}{N_h} Y_h^*. \quad (3.3)$$

Si l'on remplace (3.3) dans l'équation (3.2), on obtient $E(\hat{Y}^{st}) = Y$.

Dans le résultat principal qui suit, nous avons besoin de la notation suivante. Supposons que R_h^* et R_h' représentent respectivement les ensembles de paires non ordonnées admissibles et inadmissibles qui partent de F_h . Les définitions de ces termes sont identiques aux concepts correspondants pour l'EASSR, mais elles se limitent maintenant aux strates.

Théorème 3-2: La variance de (3.1) est:

$$V(\hat{Y}^{st}) = \sum_{h=1}^H S_h^2, \quad (3.4)$$

où,

$$S_h^2 = \frac{n_h}{N_h} \left[\sum_{k \in V_h} q_{hk} \left(\frac{Y_k}{N_h} \right)^2 + \frac{2(n_h - 1)}{(N_h - 1)} \times \right.$$

$$\left. \sum_{[h_j k, h_j' k'] \in R_h^*} \left(\frac{Y_k s_{hj_k}}{s_k} \frac{Y_{k'} s_{hj'_k}}{s_{k'}} \right) - \left(\sum_{k \in V_h} \frac{Y_k q_{hk}}{s_k} \right)^2 \right]. \quad (3.5)$$

Preuve:

Ecrivons d'abord

$$V(\hat{Y}^{st}) = E(\hat{Y}^{st})^2 - \bar{Y}^2 = E \left(\sum_{h=1}^L \hat{Y}_h^2 + 2 \left(E \left(\sum_{h < h'} \hat{Y}_h \hat{Y}_{h'} \right) - \sum_{h < h'} Y_h^* Y_{h'}^* \right) \right).$$

$$\quad (3.6)$$

Les deux derniers termes s'annulent parce que \hat{Y}_h et $Y_{h'}^*$ sont indépendants. Ceci est logique, puisque la répartition crée une nouvelle population stratifiée sans chevauchement et que les échantillons choisis dans les différentes strates sont indépendants. Par conséquent, avec

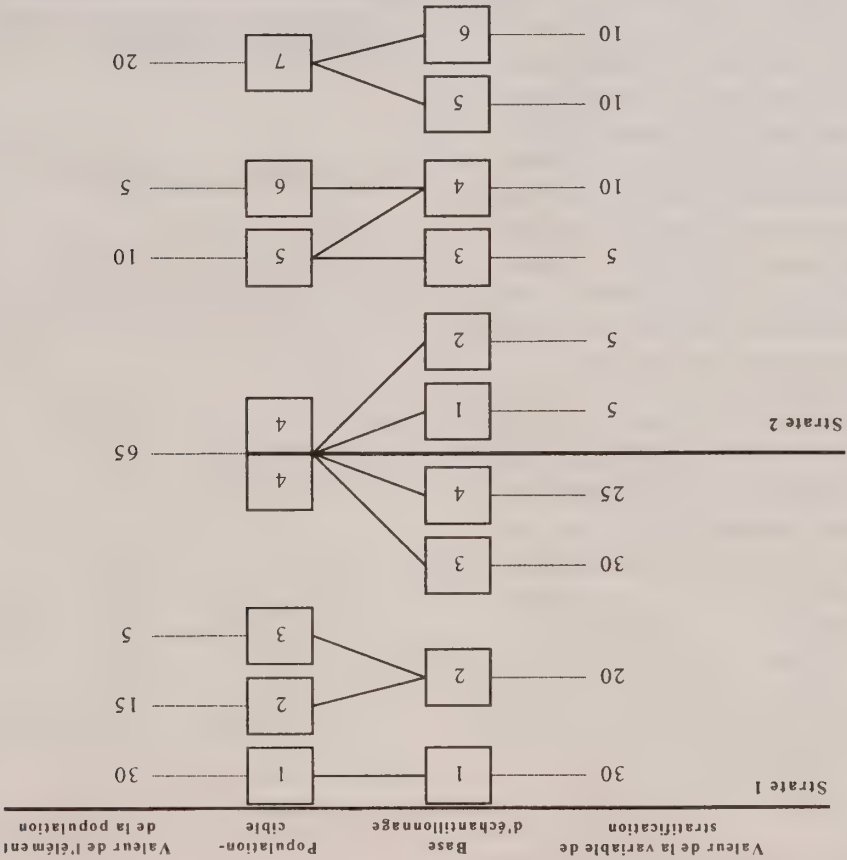


Figure 3.1. Exemple de la correspondance entre la base de sondage et la population-cible dans l'échantillonnage aléatoire stratifié

Probabilités d'arc pour la figure 3.1									
Arc hjk	111	122	123	134	144	1	1	1	s_k
	1	35796	35796	1	1	1	1	1	

Le tableau 3.2 présente les facteurs de pondération (s_k) pour tous les éléments de la population à la figure 3.1. Les observations faites à la section 2.3.1, au sujet des facteurs de pondération des arcs (s_{hjk}) et des facteurs de pondération des éléments de population (s_k), s'appliquent ici également.

Facteurs de pondération des éléments de la population (s_k) pour la figure 3.1									
k	1	2	3	4	5	6	7	s_k	

Pour chaque $h = 1, \dots, L$ et $i = 1, \dots, n_h$, supposons que x_{hk_i} est une variable aléatoire de telle sorte que x_{hk_i}/s_k si k dans T est sélectionné après la sélection de quelque unité h_j dans F_h .

où $s_{hjk} = \Pr(h_j, h_k) = k | h_j = h_j$ est une probabilité d'arc. À noter que s_{hjk} est la probabilité conditionnelle de sélectionner l'élément de population k dans T lorsque l'unité h_j de la base de sondage a été choisie. Le tableau 3.1 indique les probabilités d'arc pour la figure 3.1, lorsqu'on présume de probabilités de randomisation égales. Supposons que W_k représente l'ensemble des unités h_j dans F dont l'arc se termine à k dans T . Par exemple, $W_4 = \{(1, 3), (1, 4), (2, 1), (2, 2)\}$. Supposons également que le facteur de pondération pour l'élément de la population est $s_k = \sum_{h_j \in W_k} s_{hjk}$.

$$\Pr\{(h_j, h_k) = (h_j, h_k)\} = \frac{1}{N_h} s_{hjk}$$

l'arc aléatoire (h_j, h_k) est calculée par

1. Erreur de stratification: Par exemple, l'unité 2 de la base de sondage (adresses de voirie) dans la strate 1 semblait correspondre à un grand immeuble, en raison de la forte consommation d'électricité qui y était associée; cette unité a donc été placée dans la première strate. Les données recueillies ont toutefois révélé que l'adresse correspondait en fait à deux petits immeubles (éléments de la population 2 et 3). Dans un autre exemple, les unités 5 et 6 de la base de sondage, dans la strate 2, semblaient être deux petits immeubles et ont donc été placées dans la deuxième strate. Or l'élément de la population 7 qui y correspond est un immeuble unique ayant deux adresses.

2. Chevauchement: Par exemple, les unités 3 et 4 de la base de sondage, dans la strate 1, de même que les unités 1 et 2 dans la strate 2, ont toutes une adresse différente et figurent donc dans la base comme deux petits et deux grands immeubles. Or les données recueillies révélèrent que les quatre adresses ne correspondent en fait qu'à un seul bâtiment (par ex. un mail linéaire). Dans ce dernier cas, non seulement y a-t-il erreur de stratification, mais également les unités de la base de sondage associées à un même immeuble ne sont pas toutes incluses dans la même strate. En d'autres mots, un même élément de la population (un immeuble) chevauche plusieurs strates. Dans la section qui suit, nous élaborons des estimateurs pour le total et les chiffres de population et démontrons que ces estimateurs sont sans biais, malgré l'erreur de stratification et le chevauchement. Cependant, comme c'est habituellement le cas, l'erreur de stratification augmente la variance des estimations. En outre, dans la mesure où le chevauchement produit une erreur de stratification, celui-ci augmente également la variance des estimations.

3.3 Totaux et chiffres de population

3.3.1 Estimateur du total d'une population

L'estimateur présenté ici s'appuie sur une méthode de pondération qui consiste à prolonger l'estimateur proposé par Hansen et coll. (1953a, p. 62-64), pour l'appliquer à l'échantillonnage aléatoire stratifié avec base de sondage

plusieurs à plusieurs. Supposons que F a été divisé en L strates exhaustives s'excluant mutuellement F_1, \dots, F_L , respectivement de taille N_1, \dots, N_L . Les unités dans F_h seront représentées par h_j où $j = 1, \dots, N_h$ et $h = 1, \dots, L$. Supposons également qu'un échantillon aléatoire stratifié (sans remise) de taille $n = n_1 + \dots + n_L$ a été prélevé, où n_h est la taille de l'échantillon prélevé de F_h . Supposons que h_{j1}, \dots, h_{jn_h} représentent les variables aléatoires de sorte que $h_{ji} = h_j$ lorsque le i -ième prélevement de F_h donne lieu à la sélection de h_j . Supposons enfin que h_{K_1}, \dots, h_{K_n} représentent des variables aléatoires de sorte que $h_{K_i} = k$ si le i -ième prélevement de F_h est suivi de la sélection de k de T . Si h_j/k représente l'arc qui part de l'unité h_j dans F_h et se

Lorsqu'on introduit ces espérances dans l'équation (2.12), on obtient alors (2.11).

3. ESTIMATEURS POUR DES BASES DE STRUCTURE PLUSIEURS À PLUSIEURS, PAR ÉCHANTILLONNAGE ALÉATOIRE STRATIFIÉ

3.1 Introduction

Nous présentons dans cette section des estimateurs pour les chiffres, la moyenne et le total d'une population, avec une structure plusieurs à plusieurs selon un plan d'échantillonnage aléatoire stratifié. Il convient toutefois de décrire d'abord la méthode d'échantillonnage en vertu de laquelle ces estimations sont appropriées. La figure 3.1 présente un exemple qui sera utilisé tout au long de cette section.

3.2 Méthode d'échantillonnage

Les scénarios décrits pour l'EASSR sont également utilisés pour l'échantillonnage aléatoire stratifié. Il y a toutefois quelques problèmes additionnels qui peuvent survenir dans ce dernier cas.

Revenons à l'étude sur les caractéristiques des immeubles, qui a justifié la conduite de la présente recherche. Supposons que la taille du bâtiment est la valeur de l'élément de la population à la figure 3.1 et que la consommation d'électricité associée à l'adresse de voirie est la variable de stratification. Comme il existe un rapport de plusieurs à plusieurs entre la base de sondage (adresses de voirie) et la population-cible (immeubles commerciaux), les problèmes suivants sont venus s'ajouter à ceux mentionnés à la section 2.1 :

devient, avec un plan d'EASSR et une structure de plusieurs à plusieurs :

$$\bar{Y}_{\frac{\Delta}{2}} = \frac{\sum_{i=1}^n x_{k_i}}{\sum_{i=1}^n z_{k_i}}. \quad (2.10)$$

2.2.2 Erreur quadratique moyenne (EQM) de la moyenne d'une population

L'estimateur de la moyenne d'une population est biaisé, parce qu'il s'agit d'un estimateur par quotient. Cependant, on sait très bien que ce biais devient négligeable avec de gros échantillons et que le biais est d'ordre $1/n$ (Cochran 1977, p. 160).

Notre approximation de l'erreur quadratique moyenne exige une sommation de $R^{**} - 1$ ensemble de toutes les paires d'arcs *ordonnées admissibles*. Par conséquent, si $[jk, j'k'] \in R^*$, alors $[j'k', jk] \in R^{**}$.

Pour calculer la valeur approximative de l'erreur quadratique moyenne de l'estimateur (2.10), nous utilisons

$$EQM(\bar{Y}_{\frac{\Delta}{2}}) \approx \frac{nN \left(\sum_{k=1}^M \frac{1}{z_{k_i}} \right)^2}{M^2}$$

$$\left[\left(\sum_{k=1}^M \frac{y_k}{z_k} + \frac{2(N-1)}{2} \sum_{[jk, j'k'] \in R^*} \frac{y_k s_{jk}}{y_{k'} s_{j'k'}} \right) \right]$$

$$- 2\bar{Y} \left(\sum_{k=1}^M \frac{y_k}{z_k} + \frac{(N-1)}{2} \sum_{[jk, j'k'] \in R^{**}} \frac{y_k s_{jk}}{y_{k'} s_{j'k'}} \right) \left[\right]$$

$$+ \bar{Y}^2 \left(\sum_{k=1}^M \frac{1}{z_k} + \frac{(N-1)}{2} \sum_{[jk, j'k'] \in R^*} \frac{s_k}{s_{j'k'}} \right) \left[\right]. \quad (2.11)$$

Pour justifier cette approximation, supposons que

$$\bar{x} = \frac{\sum_{i=1}^n x_{k_i}}{\sum_{i=1}^n z_{k_i}}, \quad \bar{z} = \frac{n}{\sum_{k=1}^M \frac{1}{z_{k_i}}}, \quad \text{et } \bar{Z} = \frac{M}{\sum_{k=1}^M \frac{1}{z_{k_i}}}.$$

Comme \bar{Y} est un ratio de deux estimations, l'approximation bien connue de l'erreur quadratique moyenne (Cochran 1977, p. 32-33) peut être utilisée:

$$\begin{aligned} & E \left(\sum_{n=1}^n \sum_{i=1}^{i \neq i'} x_{k_i} \frac{1}{s_{k_i}} \right) = n(n-1) E \left(x_{k_i} \frac{1}{s_{k_i}} \right) \\ & = n(n-1) \Pr \left(x_{k_i} = \frac{y_k}{z_k}, \frac{1}{s_{k_i}} = \frac{1}{s_{j'k'}} \right) \\ & = \sum_{[jk, j'k'] \in R^{**}} \sum_{[jk, j'k'] \in R^{**}} \frac{y_k}{z_k} \frac{1}{s_{k_i}} \left(\frac{N(N-1)}{1} \right) \\ & = \sum_{[jk, j'k'] \in R^{**}} \sum_{[jk, j'k'] \in R^{**}} \frac{y_k s_{jk}}{y_{k'} s_{j'k'}}. \end{aligned}$$

Si l'on utilise (2.7) et (2.9), on obtient

$$\begin{aligned} & E \left(\sum_{n=1}^n \sum_{i=1}^{i \neq i'} x_{k_i} \frac{1}{s_{k_i}} \right) = \frac{n}{N} \sum_{k=1}^M \frac{y_k}{z_k} + E \left(\sum_{n=1}^n \sum_{i=1}^{i \neq i'} x_{k_i} \frac{1}{s_{k_i}} \right) \\ & = E \left(\sum_{n=1}^n \sum_{i=1}^{i \neq i'} z_{k_i} \right) = E \left(\sum_{n=1}^n x_{k_i} \frac{1}{s_{k_i}} \right) + \end{aligned}$$

La première espérance dans (2.12) est tout simplement (2.9). Ensuite, l'utilisation de l'équation (2.1) dans le terme central de (2.12) donne

$$\begin{aligned} & \frac{1}{N} \left[E(\bar{x}^2) - 2\bar{Y}E(\bar{z}\bar{x}) + \bar{Y}^2 E(\bar{z}^2) \right] = \\ & \frac{1}{N} \left[E(\bar{x}^2) - 2\bar{Y}E(\bar{z}\bar{x}) + \bar{Y}^2 E(\bar{z}^2) \right] = \end{aligned} \quad (2.12)$$

$$\frac{1}{N} \left[E(\bar{x}^2) - 2\bar{Y}E(\bar{z}\bar{x}) + \bar{Y}^2 E(\bar{z}^2) \right] =$$

$$EQM(\bar{Y}_{\frac{\Delta}{2}}) = E \left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{x} - \bar{Y}\bar{z}} \right)^2 \approx E \left(\frac{\bar{z}}{\bar{x} - \bar{Y}\bar{z}} \right)^2 =$$

À noter que la double somme englobe toutes les paires *ordonnées* et admissibles d'arcs. Par conséquent,

Compte tenu de l'indépendance de la randomisation et du choix des unités de la base de sondage:

$\Pr(\text{sélection } [j, k, j', k'] \text{ dans } R^*) = \Pr(\text{sélection } \{j, j'\})$
dans $F)$

$\Pr(\text{sélection } [j, k, j', k'] \text{ dans } R^* \mid \text{sélection } \{j, j'\})$ dans

$$F) = \frac{1}{N} \binom{2}{2} s_{jk} s_{j'k'}$$

Si l'on substitue ceci dans (2.7), on obtient alors

$$n(n-1) \sum \sum_{[jk, j', k'] \in R^*} \left[(x_k x_{k'}) \frac{1}{N} \binom{2}{2} s_{jk} s_{j'k'} \right] =$$

$$\frac{2n(n-1)}{N(N-1)} \sum \sum_{[jk, j', k'] \in R^*} \left[\left(\frac{s_k}{y_k s_{jk}} \frac{s_{k'}}{y_{k'} s_{j'k'}} \right) \right]. \quad (2.8)$$

Si l'on introduit maintenant (2.6) et (2.8) dans (2.5), on obtient,

$$E \left[\sum_{i=1}^n x_{k_i} \right]^2 = \frac{n}{N} \sum_{k=1}^M \frac{y_k}{s_k} +$$

$$\frac{2(n-1)}{N(N-1)} \sum \sum_{[jk, j', k'] \in R^*} \left[\left(\frac{s_k}{y_k s_{jk}} \frac{s_{k'}}{y_{k'} s_{j'k'}} \right) \right]. \quad (2.9)$$

Enfin, si l'on introduit (2.9) dans (2.4), nous obtenons le

résultat en (2.3).

L'équation (2.3) est une généralisation de la formule élaborée par Bandyopadhyay et Adhikari (1993) pour calculer la variance de l'estimation du total d'une population, avec une structure à plusieurs. On peut voir que l'équation (2.3) correspond à leur formule, lorsque la base de sondage est limitée à une structure à plusieurs à un.

Corollaire 2-1: Voici une autre façon d'exprimer la formule de la variance dans le **Théorème 2-2:**

$$V(\hat{Y}) = \frac{n}{N} \left[\sum_{k=1}^M \frac{y_k}{s_k} + \frac{(N-1)}{(n-1)} \left(\sum_{jk} \frac{y_k}{s_k} \frac{s_k}{s_{jk}} \right)^2 - \sum_{jk} \sum_{[jk, j', k'] \in R^*} \frac{y_k s_{jk}}{y_{k'} s_{j'k'}} \frac{s_k}{s_{k'}} \right] - Y^2.$$

2.2 Moyenne d'une population

2.2.1 Estimateur de la moyenne d'une population

L'estimateur de la moyenne d'une population, qui est présenté ici, est un prolongement de l'estimateur présenté par Hansen et coll. (1953a), ici appliqué à la structure à plusieurs à plusieurs.

En rapport avec les n prélèvements de F , définissons les variables aléatoires s_{k_i} et $z_{k_i} = 1/s_{k_i}$, de sorte que s_{k_i} a la valeur s_k si $K_i = k$ pour $i = 1, \dots, n$ et $k = 1, \dots, M$. L'estimateur de la moyenne d'une population,

$$\bar{Y} = \frac{1}{M} \sum_{k=1}^M y_k,$$

Nous obtenons le résultat en remplaçant l'expression qui précède dans (2.3).

Cette formule est plus simple au plan computationnel. À noter que l'équation (2.3) exige que l'on fasse la somme du terme

$$\left(\frac{s_k}{y_k s_{jk}} \frac{s_{k'}}{y_{k'} s_{j'k'}} \right)$$

Il s'ensuit que:

$$\sum \sum_{[jk, j', k'] \in R^*} \frac{s_k}{y_k s_{jk}} \frac{s_{k'}}{y_{k'} s_{j'k'}} = \frac{1}{2} \left(\sum_{jk} \frac{s_k}{y_k s_{jk}} \right)^2 - \sum \sum_{[jk, j', k'] \in R^*} \frac{s_k}{y_k s_{jk}} \frac{s_{k'}}{y_{k'} s_{j'k'}}.$$

$$\left(\sum_{jk} \frac{y_k}{s_k} \frac{s_k}{s_{jk}} \right)^2 = \sum_{jk} \left(\frac{s_k}{y_k s_{jk}} \right)^2 + 2 \sum \sum_{[jk, j', k'] \in R^*} \frac{s_k}{y_k s_{jk}} \frac{s_{k'}}{y_{k'} s_{j'k'}}.$$

Preuve: Si nous écrivons

inadmissibles d'arcs. L'ensemble des paires non ordonnées $admissibles$ d'arcs est la série complémentaire $R^* = P \setminus R'$. Pour illustrer ceci, examinons la figure 2.1. En vertu de la méthode d'échantillonnage utilisée, si l'unité 4 de la base de sondage est sélectionnée, un seul des deux éléments de population 3 ou 4 peut être inclus dans l'échantillon. Par conséquent, $\{[4,3][4,4]\}$ est une paire non ordonnée inadmissible d'arcs. Les autres paires d'arcs non ordonnées et inadmissibles à la figure 2.1 sont $\{[6,5][6,6]\}$ et $\{[7,5][7,6]\}$. Par conséquent, $R' = \{[4,3][4,4], [6,5][6,6], [7,5][7,6]\}$.

Théorème 2-2: La variance de l'estimateur (2.2) est

$$V(\hat{Y}) = \frac{n}{N} \left[\sum_{k=1}^M \frac{Y_k^2}{s_k} + 2 \frac{(N-1)}{(n-1)} \sum \right]$$

$$\sum_{[jk,j'k'] \in R^*} \left(\frac{Y_k s_{jk}}{Y_{k'} s_{j'k'}} \right) - Y^2, \quad (2.3)$$

où la double somme englobe toutes les paires non ordonnées *admissibles* d'arcs $[jk, j'k']$.

Preuve:

$$V(\hat{Y}) = E \left[\left(\frac{n}{N} \sum_{k=1}^M x_{k_i} \right)^2 \right] - Y^2$$

$$= \frac{n^2}{N^2} E \left[\left(\sum_{i=1}^n x_{k_i} \right)^2 \right] - Y^2. \quad (2.4)$$

Maintenant,

$$E \left[\left(\sum_{i=1}^n x_{k_i} \right)^2 \right] = \sum_{i=1}^n E(x_{k_i}^2) + 2E \left(\sum_{i < j} x_{k_i} x_{k_j} \right). \quad (2.5)$$

On peut écrire

$$E(x_{k_i}^2) = \sum_{k=1}^M x_k^2 \Pr(K_i = k) = \sum_{k=1}^M \frac{Y_k^2}{s_k} \frac{s_k}{N} = \frac{1}{M} \sum_{k=1}^M \frac{Y_k^2}{s_k}. \quad (2.6)$$

Comme nous l'avons indiqué à la section 2.1, nous pouvons sélectionner un échantillon d'arcs qui mène ensuite à la sélection des éléments de la population. Chaque arc (jk) est associé à une valeur $x_k = Y_k/s_k$ de l'élément de la population k à sa destination. Nous pouvons donc récrire la double sommation en (2.5) comme étant le cumul des paires non ordonnées admissibles d'arcs, R^* .

$$2E \left(\sum_{i < j} x_{k_i} x_{k_j} \right) = 2 \left(\sum_{[jk,j'k'] \in R^*} \sum \right) \left(x_k x_{k'} \Pr(K_i = k, K_{j'} = k') \right). \quad (2.7)$$

Maintenant, supposons que x_1, \dots, x_M représentent les valeurs pondérées associées aux indices dans T , c'est-à-dire supposons que $x_k = Y_k/s_k$. Définissons les variables aléatoires x_{k_1}, \dots, x_{k_n} , associées respectivement aux prélèvements 1 à n à partir de F , de sorte que x_{k_i} prend la valeur x_k si $K_i = k$. Nous pouvons alors écrire

$$E(x_{k_i}) = \sum_{k=1}^M x_k \Pr(K_i = k) = \frac{1}{M} \sum_{k=1}^M \frac{Y_k}{s_k} s_k = \frac{Y}{N}, \quad (2.1)$$

où $Y = \sum_{k=1}^M Y_k$ est le véritable total de la population. Nous utilisons comme estimateur du total de la population, selon un plan FASSR avec structure à plusieurs,

$$\hat{Y} = \frac{n}{N} \sum_{i=1}^n x_{k_i}. \quad (2.2)$$

Si l'on utilise (2.1), il s'ensuit que

$$E(\hat{Y}) = E \left(\frac{n}{N} \sum_{i=1}^n x_{k_i} \right) = \frac{n}{N} \sum_{i=1}^n E(x_{k_i}) = \frac{n}{N} \frac{Y}{N} = Y.$$

Nous obtenons alors

Théorème 2-1: L'estimateur (2.2) du total d'une population, utilisé avec un plan d'FASSR, est sans biais.

À partir de la figure 2.1, nous présentons maintenant un exemple simple de l'utilisation de cet estimateur. Supposons qu'un échantillon aléatoire simple formé de quatre unités a été sélectionné de la base de sondage illustrée par la figure 2.1 (2, 3, 4 et 7), ce qui a mené par la suite à la sélection des éléments de population 2, 4 et 5. L'estimateur du total de la population,

$$\hat{Y} = \frac{n}{N} \sum_{i=1}^n x_{k_i}, \text{ à la valeur } \frac{4}{385} \left[\frac{7}{20} + \frac{2}{20} + \frac{(1/2)}{15} + \frac{2}{10} \right] = \frac{4}{385}.$$

L'estimateur qui précède peut également être utilisé pour obtenir les chiffres de population. Nous pourrions ainsi estimer la taille de la population-cible en supposant que $Y_k = 1$ pour tous les k . Nous pourrions également estimer le nombre d'éléments de la population qui possèdent certaines caractéristiques, en supposant que $Y_k = 1$ pour les éléments de la population qui présentent les caractéristiques qui nous intéressent et que $Y_k = 0$ pour ceux qui n'ont pas ces caractéristiques.

2.1.2 Variance de l'estimateur du total d'une population

Il convient en premier lieu de définir certains autres termes et notations utilisés dans cette section. Supposons que P représente l'ensemble de toutes les paires non ordonnées d'arcs. Une paire non ordonnée d'arcs est dite *inadmissible* si les deux éléments ne peuvent être inclus dans un échantillon. Supposons que $\mathcal{O} = \{j \text{ dans } F; \text{ plus d'un arc émerge de } j\}$. Alors $R' = \{[jk, j'k']; j \in \mathcal{O} \text{ et } k \neq k'\}$ représente l'ensemble des paires non ordonnées

Dans le deuxième scénario, plusieurs unités de la base de sondage correspondent à un élément de la population (structure plusieurs à un). À la figure 2.1, les unités 2 et 3 de la base de sondage correspondent à un seul élément de la population, le 2. Dans ce dernier cas, si les unités 2 ou 3, ou les deux, de la base de sondage sont incluses dans l'échantillon, on obtient alors de l'information sur l'élément 2 de la population. Il est donc possible que cet élément de la population (2) apparaisse jusqu'à deux fois dans l'échantillon – et aussi comme enregistré, dans le fichier de données utilisé pour faire les estimations.

Dans le troisième scénario, une unité de la base de sondage correspond à plus d'un élément de la population (structure un à plusieurs). Toujours à la figure 2.1, l'unité 4 correspond aux éléments de la population 3 et 4. Ici, un seul élément de la population (3 ou 4) est choisi par une méthode de *randomisation* indépendante du choix des unités de la base de sondage. Cette méthode a été justifiée par des motifs économiques, car la collecte des données nécessitait de longues interviews sur place menées par des personnes ayant une formation technique. Nous présumons ici que les probabilités de *randomisation* sont égales. Cependant, toute autre probabilité différente de zéro pourrait également être utilisée (par ex. probabilité proportionnelle à la taille).

Le quatrième scénario prévoit une structure de plusieurs à plusieurs, laquelle est illustrée par les unités 5, 6 et 7 de la base de sondage et les éléments 5 et 6 de la population, à la figure 2.1. Comme ces cas complexes sont en fait une combinaison des scénarios 2 et 3 qui précèdent, les mêmes règles d'échantillonnage s'appliquent. Ainsi, si l'unité 5 est choisie, c'est l'élément de la population 5 qui est mesuré. Si l'unité 6 est choisie, un seul des deux éléments de la population 5 ou 6 est choisi au hasard et mesuré.

2.1 Total d'une population

2.1.1 Estimateur du total d'une population

Une base de sondage de plusieurs à plusieurs donne lieu à diverses probabilités de sélection. Les estimateurs proposés ici sont basés sur une méthode de pondération et constituent un prolongement de ceux présentés par Hansen et coll. (1953a p. 62-64). Les estimateurs proposés par ces auteurs, de même que les formules pour en calculer la variance, se limitent à la structure plusieurs à un; nous avons prolongé ces estimateurs pour les appliquer à une structure plusieurs à plusieurs.

Dans un plan d'EASSR d'effectifs n , supposons que J_1, \dots, J_n représentent des variables aléatoires de telle sorte que $J_i = j$ si le i -ième prélèvement mène à la sélection de l'unité j dans F . Donc $\Pr(J_i = j) = 1/N$ pour j dans F et $i = 1, \dots, n$. Supposons maintenant que K_1, \dots, K_n représentent des variables aléatoires de sorte que $K_i = k$ si le i -ième prélèvement dans F est suivi de la sélection de k dans T . Nous pouvons maintenant passer au prélèvement d'un échantillon aléatoire d'arcs $\{(J_1 K_1), \dots, (J_n K_n)\}$ dont la distribution de probabilité conjointe est déterminée à la fois par le plan d'échantillonnage EASSR et la *randomisation*

subséquent (s'il y a lieu) pour choisir un élément dans T . $(J_i K_i)$ a une probabilité marginale représentée par $\Pr\{(J_i K_i) = (jk)\} = (1/N)s_{jk}$, où $s_{jk} = \Pr(K_i = j | J_i = j)$. En d'autres mots, s_{jk} est la probabilité conditionnelle de sélectionner l'élément de population k dans T lorsque l'unité de la base de sondage j dans F est choisie. Ces probabilités conditionnelles, désignées probabilités d'arc, sont illustrées pour la figure 2.1 au tableau 2.1.

Tableau 2.1

Probabilités d'arc pour la figure 2.1

Arc j/k	1,1	2,2	3,2	4,3	4,4	5,5	6,5	7,5	7,6
s_{jk}	1	1	1	1/2	1/2	1	1/2	1/2	1/2

Pour k dans T , supposons que U_k représente l'ensemble des unités dans F dont les arcs mènent à k dans T . Supposons également que $s_k = \sum_{j \in U_k} s_{jk}$. Si nous reprenons le langage de Hansen et coll. (1953a p. 62-64), sur lequel s'appuie notre raisonnement, s_k est le *facteur de pondération* de l'élément de la population k dans T . Les facteurs de pondération pour la figure 2.1 sont indiqués au tableau 2.2.

Tableau 2.2

Calcul des facteurs de pondération (s_k) des éléments de la population pour la figure 2.1

k	1	2	3	4	5	6
(s_k)	1	2	1/2	1/2	2	1

Les probabilités d'arc et les facteurs de pondération sont utilisés pour calculer les probabilités marginales de K_i , nommée $\Pr(K_i = k) = \sum_{j \in U_k} (1/N)s_{jk} = (1/N)s_k$, à savoir lorsque k est dans T et $i = 1, \dots, n$. De toute évidence, le calcul des probabilités d'arc est l'étape cruciale dans l'élaboration des facteurs de pondération appropriés pour les données recueillies. Ce calcul dépend de la détermination adéquate de la structure de graphe pour chaque unité d'échantillonnage choisie: dans un sous-graphe connexe maximal (CM). Un sous-graphe connexe désigne un sous-ensemble de noeuds reliés par une série d'arcs. Maximal signifie qu'aucun noeud à l'extérieur du sous-ensemble n'est relié à un noeud qui appartient au sous-ensemble. Il y a 4 sous-graphes CM à la figure 2.1; chacun représente une structure base – population différente, à savoir les structures un à un, plusieurs à un, un à plusieurs et plusieurs à plusieurs. Il n'est pas nécessaire de connaître la structure du graphe entier pour définir les estimateurs. Il suffit seulement de connaître la structure des sous-graphes CM auxquels appartiennent les unités *échantillonnées* de la base.

Nous faisons les observations suivantes au sujet de s_k et s_{jk} : i) $s_k = W$ indique que la probabilité de sélectionner l'élément de population k au i -ième prélèvement est de W fois celle d'un élément dont le facteur de pondération est un; ii) $0 < s_k \leq N$, $k = 1, \dots, M$; iii) $0 < s_{jk} \leq 1$, $j \in U_k$ et $k = 1, \dots, M$; iv) avec la structure un à plusieurs, $s_{jk} = s_k$; v) avec la structure plusieurs à un, $s_{jk} = 1$ pour tous les k et vi) $\sum_{k=1}^M \sum_{j=1}^N s_{jk} = N$.

d'une base de sondage de structure plusieurs à plusieurs. Mentionnons d'abord l'estimateur de Horvitz-Thompson, (1952), qui fournit des estimations sans biais de la moyenne et du total d'une population lorsqu'il existe plusieurs probabilités de sélection. Musser (1993) montre comment calculer les bonnes probabilités d'inclusion pour les éléments de la population sélectionnés par échantillonnage aléatoire simple à partir d'une base de plusieurs à un. Cependant, la méthode de Musser peut également être appliquée au calcul des probabilités d'inclusion pour des éléments de population dans un échantillon aléatoire simple prélevé d'une base de sondage de structure plusieurs à plusieurs. En deuxième lieu, Lavallée (1995) a adapté la méthode à poids partagés – utilisée pour les enquêtes longitudinales – pour l'appliquer aux bases de sondage plusieurs à plusieurs.

Le but du présent article est de proposer une autre méthode pour estimer le total, les chiffres et la moyenne de la population, avec des bases de sondage de structure plusieurs à plusieurs, selon des plans d'échantillonnage aléatoire simple et stratifié. Nous calculons également les expressions de la variance de ces estimateurs. Les résultats que nous présentons n'ont pas seulement un intérêt intrinsèque; en effet, les expressions de la variance des estimateurs sont essentielles à l'étude des effets des imperfections de correspondance inhérentes aux bases de sondage plusieurs à plusieurs sur la précision de ces estimations.

Nous présentons à la section 2 les estimations obtenues par échantillonnage aléatoire simple sans remise (EASSR). Nous y décrivons également la méthode d'échantillonnage avec laquelle ces estimateurs sont applicables, puis présentons un résultat du biais et proposons des façons d'en exprimer la variance.

À la section 3, certains de ces résultats sont appliqués à l'échantillonnage aléatoire stratifié. Enfin, à la section 4, nous tirons quelques conclusions, discutons des limites de la méthode et proposons des suggestions pour de futurs projets de recherche.

2. BASES DE SONDAJE PLUSIEURS À PLUSIEURS POUR L'ÉCHANTILLONNAGE ALÉATOIRE SIMPLE

Il est utile de représenter sous forme de graphique la relation entre la base de sondage et la population-cible. Les unités d'échantillonnage dans la base de sondage et les éléments de la population-cible forment les deux ensembles de noeuds; des arcs relient les unités d'échantillonnage aux éléments de la population-cible. Ces arcs illustrent la structure du rapport entre la base de sondage et la population-cible. La figure 2.1 présente un exemple d'une base de sondage et d'une population-cible avec un rapport de plusieurs à plusieurs. Il y a 7 unités d'échantillonnage dans la base, 6 éléments dans la population-cible et 10 liens (arcs) entre les unités d'échantillonnage et les éléments de la population. Cette structure de plusieurs à plusieurs est

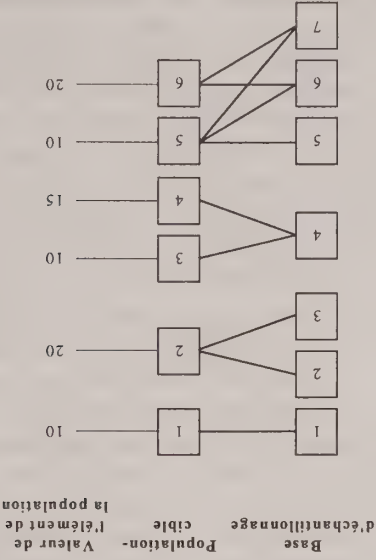


Figure 2.1

Exemple de la correspondance entre la base de sondage et la population-cible

donc représentée par un graphique formé de 13 noeuds et de 10 arcs. Dans la présente analyse, nous supposons que chaque élément de la population est relié aux unités de la base par au moins un arc et que chaque unité de la base est liée aux éléments de la population, la aussi par au moins un arc.

Établissons maintenant quelques notations. Il nous paraît commode d'identifier les unités de la base et les éléments de la population par leurs indices respectifs. Supposons que $F = \{1, 2, \dots, N\}$ représente l'ensemble des indices pour N unités d'échantillonnage et que $T = \{1, 2, \dots, M\}$ représente l'ensemble des indices pour les M éléments de la population-cible. Un arc peut être représenté comme étant une paire ordonnée, dont le premier élément provient de F et le deuxième, de T . On dit qu'un élément de la population k dans T est représenté par l'unité d'échantillonnage j dans F s'il y est relié par un arc désigné (jk) . Ceci signifie que, lorsque j est dans l'échantillon, il existe une probabilité différente de zéro de recueillir des données à partir de l'élément de la population k . Nous représenterons par y_k la mesure de l'élément de la population k dans T qui nous intéresse.

Nous décrivons maintenant la technique d'échantillonnage en vertu de laquelle les estimateurs proposés ici peuvent s'appliquer. Supposons que n unités de la base sont sélectionnées dans F , par EASSR. Le nombre d'éléments de la population inclus dans l'échantillon et mesurés dépend toutefois de la nature du lien qui existe entre les unités de la base de sondage et les éléments de la population.

Dans le cas de l'EASSR, quatre scénarios sont possibles lorsqu'une unité est sélectionnée. Dans le premier scénario, l'unité de la base ne correspond qu'à un seul élément de la population (structure un à un). Ici, l'enquêteur recueillera des données uniquement sur l'élément de la population qui correspond à l'unité sélectionnée de la base (voir unité 1 de la base de sondage, à la figure 2.1).

Estimations à partir de bases de sondage de structure plusieurs à plusieurs

FERRI L. BYCZKOWSKI, MARTIN S. LEVY et DENNIS J. SWEENEY¹

RÉSUMÉ

Pour les sondages, il doit idéalement y avoir correspondance de un à un entre les unités de la base de sondage et les éléments de la population-cible à l'étude. Dans bien des cas, toutefois, la base de sondage a une structure plusieurs à plusieurs, c'est-à-dire qu'une unité de la base de sondage peut être associée à de multiples éléments de la population-cible et, vice-versa, un élément de la population-cible peut être associé à de multiples unités de la base de sondage. C'est ce qui s'est produit dans le cadre d'une enquête sur les caractéristiques des immeubles, pour laquelle la base de sondage était constituée des adresses de voirie et les immeubles commerciaux formaient la population-cible. La base de sondage était complexe, car une même adresse pouvait correspondre à un seul immeuble, à plusieurs immeubles ou à une partie d'un immeuble. Nous présentons ici des estimateurs et des formules pour calculer la variance, selon des plans d'échantillonnage aléatoire simple et stratifié, lorsque la base de sondage est de structure plusieurs à plusieurs.

MOTS-CLÉS : Bases de sondage imparfaites; erreurs de correspondance; enquête sur les caractéristiques des immeubles; pondération; échantillonnage aléatoire simple; échantillonnage aléatoire stratifié.

1. INTRODUCTION

Cette recherche fait suite à une étude qui avait été menée pour le compte d'une société de services publics, dans le but d'estimer diverses caractéristiques des immeubles commerciaux (population-cible) situés dans la région desservie par cette société. La liste des immeubles commerciaux ne pouvait être établie par dénombrement sur le terrain, en raison des coûts qu'aurait engendrés une telle méthode. On disposait toutefois d'une base de sondage constituée des adresses de voirie (c.-à-d. des adresses où il y avait un compte), dont une des lacunes venait du rapport de plusieurs à plusieurs entre la base et la population-cible (immeubles commerciaux); ainsi, certaines unités de la base de sondage étaient associées à de multiples éléments de la population-cible et certains éléments de la population-cible étaient associés à un grand nombre d'unités de la base de sondage. En fait, plusieurs rapports entre les adresses et les immeubles étaient relativement complexes.

Cependant, un des avantages de cette base était qu'elle fournissait la consommation annuelle totale d'électricité pour chaque adresse de voirie; nous disposions donc d'une variable permettant la stratification efficace de la base des adresses. La superficie commerciale totale était une des caractéristiques importantes à mesurer; des études menées aux États-Unis ont en effet révélé que la consommation d'énergie est fonction à la fois de la taille du bâtiment et de l'activité qui y est menée. À titre d'exemple, la consommation est plus élevée dans les immeubles utilisés pour la prestation de soins de santé ou la vente d'aliments, alors qu'elle est plus faible dans les immeubles utilisés pour des

cérémonies religieuses ou des assemblées publiques. Il existe également une corrélation entre la consommation d'énergie et la taille du bâtiment, même lorsque l'utilité du bâtiment n'est pas connue, comme c'était le cas ici (U.S. Department of Energy 1992).

Les bases de sondage imparfaites ont fait l'objet de nombreuses études, dont on peut en trouver des résumés détaillés dans Kish (1965), Wright et Tsao (1983) et Lessler et Kalisbeek (1992). D'autres ouvrages traitent de l'échantillonnage par multiplicité, où la base de sondage est conçue de manière à avoir une structure de plusieurs à plusieurs. Dans ce dernier cas, des imperfections sont introduites dans la base de sondage, afin de recueillir plus efficacement de l'information sur les occurrences qui sont rares dans une population (Birnbaum et Sirken 1965, Sirken 1972a,b et Casady et Sirken 1980). Hansen, Hurwitz et Madow (1953a,b) proposent un estimateur à utiliser avec les bases de sondage qui ont une structure plusieurs à un, c'est-à-dire où les éléments de la population sont représentés plusieurs fois dans la base. Cet estimateur est celui qu'a choisi le National Agricultural Statistics Service (NASS) pour ses enquêtes (Mussler 1993), lorsque la base a une structure plusieurs à un. Bandyopadhyay et Adhikari (1993) proposent quant à eux des estimateurs pour un quotient, la moyenne d'une population et le total d'une population, lorsque le degré de répétition dans la base de sondage est inconnu. Cependant, ces estimateurs ne peuvent être utilisés que dans les cas d'échantillonnage aléatoire simple avec base de sondage de plusieurs à un.

La documentation propose également deux méthodes pour estimer les caractéristiques de la population à partir

BIBLIOGRAPHIE

- ARMSTRONG, J., et ST-JEAN, H. (1994). Estimation par régression généralisée pour un échantillon à deux phases de dossiers fiscaux. *Techniques d'enquête*, 20, 101-110.
- BINDER, D.A. (1996). Méthodes de linéarisation pour les échantillons à une et deux phases: Une approche de type «recette». *Techniques d'enquête*, 22, 17-22.
- BINDER, D.A., BABYAK, C., BRODEUR, M., HIDIROGLOU, M.A., et JOCELYN, W. (1997). Variance Estimation for Two-phase Stratified Sampling. Article présentée à l' Annual Meeting of the American Statistical Association, Los Angeles.
- BREIDT, J., et FULLER, W.A. (1993). Regression weighting for multistage samples. *Sankhyā*, 55, 297-309.
- CHAUDHURI, A., et ROY, D. (1994). Model assisted survey sampling strategy in two phases. *Metrika*, 41, 355-362.
- COCHRAN, W.G. (1977). *Sampling Techniques*, (3-ième éd.). New York: John Wiley.
- DEVILLE, J.-C., et SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.-C., SÄRNDAL, C.-E., et SAUTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- DUPONT, F. (1995). Redressements alternatifs en présence de plusieurs niveaux d'information auxiliaire. *Techniques d'enquête*, 21, 141-150.
- ESTEVAO, V., HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, 11, 181-204.
- HIDIROGLOU, M.A. (1995). Sampling and estimation for stage one of the Canadian Survey of Employment, Payrolls and Hours survey redesign. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 123-128.
- HIDIROGLOU, M.A., LATOUCHE, M., ARMSTRONG, B., et GOSSEN, M. (1995). Improving survey information using administrative records: the case of the canadian employment survey. *Proceedings of the 1995 Annual Research Conference*. U.S. Bureau of the Census, 171-197.
- HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1995). Use of auxiliary information for two-phase sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 873-878.
- NEYMAN, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101-116.
- SÄRNDAL, C.-E., et SWENSSON, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *Revue Internationale de Statistique*, 55, 279-294.
- SÄRNDAL, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New-York: Springer-Verlag.
- SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateur de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.
- STUKEL, D., HIDIROGLOU, M.A., et SÄRNDAL, C.-E. (1996). Estimation de la variance des estimateurs de calage: comparaison des méthodes du jackknife et de la linéarisation de Taylor. *Techniques d'enquête*, 22, 117-126.

Il s'ensuit qu'on peut écrire l'estimateur (6.1) sous la forme $\hat{Y}(d) = \sum_{i=1}^I \hat{Y}_{ij}(d)$ avec

$$\hat{Y}_{ij}(d) = G_{1i} \hat{N}_{1ij} \{ \bar{y}_{s_{2ij}}(d) + (\bar{\zeta}_{s_{1ij}} - \bar{\zeta}_{s_{2ij}})' \mathbf{B}_j(d) \}$$

où

$$\bar{y}_{s_{2ij}}(d) = \sum_{k \in s_{2ij}} w_k^* y_k(d) / N_{2ij}$$

et $\hat{N}_{2ij} = \sum_{k \in s_{2ij}} w_k^*$.

Les plans d'échantillonnage à deux phases présentent l'avantage d'être économiques et efficaces. Cet article propose une théorie générale, applicable à de tels plans quand on dispose de données auxiliaires à chaque étape. L'objectif consiste à intégrer cette méthode des enquêtes à deux phases au Système généralisé d'estimation (SGE) de Statistique Canada décrit dans Estéva et coll. (1995). Le SGE est un programme général actuellement utilisé pour l'estimation de domaines dans le contexte de plans d'échantillonnage à une phase arbitraires. Il incorpore des données auxiliaires au processus d'estimation. L'article élargit les principes fondamentaux du SGE aux plans d'échantillonnage à deux phases, y compris l'idée majeure des groupes de calage.

Enfin, nous avons illustré l'application de cette théorie grâce à deux enquêtes courantes de Statistique Canada. De par sa nature générale, la théorie pourrait s'appliquer à tout plan d'échantillonnage à deux phases qui fait appel à des données auxiliaires.

8. CONCLUSIONS

pour toutes les valeurs $k \in s_{2ij}$. Ici N_{ij} est défini dans (7.1), tandis que N_{1ij} et N_{2ij} sont les mêmes que dans (7.3). Ces facteurs de calage globaux viennent du produit de deux facteurs de calage obtenus par stratification a posteriori. Ils sont positifs et bien définis, pourvu que les cellules de l'échantillon s_{2ij} ne soient pas vides. On préconise le groupement des petites cellules de s_{2ij} à des cellules non vides de plus grande taille pour une estimation stable. Ainsi qu'on le soulignait dans la remarque 3.4, les poids globaux dérivant de (7.5) reproduisent la taille connue N_{ij} des strates a posteriori de la première phase, mais ceux issus de l'équation (7.2) ne le font pas.

Remarque 7.1: Comparons les facteurs de calage (7.2) et (7.5) qui découlent respectivement de la forme réduite et de la forme complète (7.4). Ces deux facteurs $x_k = \Delta_{2k}$ sont le produit de deux termes. La seule distinction se situe au niveau du second terme. Dans les deux cas, le calcul du deuxième terme nécessite des données recoupées. En d'autres mots, pour chaque valeur $k \in s_1$, il faut identifier la cellule de recoupement ij à laquelle appartient k . En ce qui concerne le vecteur réduit, l'information des cellules est regroupée dans les groupes de la première phase. L'information reste toutefois distincte pour le vecteur complet, de sorte que les poids résultants devraient s'avérer plus efficaces.

Remarque 7.2: Pour le deuxième calage, une solution de rechange à (7.4), qui saisit aussi l'information relative aux strates a posteriori de la première phase, consiste à prendre Notons qu'avec une telle spécification, on ne compte qu'un groupe de calage à la deuxième phase, en l'occurrence tout l'échantillon s_1 de la première phase.

7.2 Cas de l'Enquête canadienne sur l'emploi, la rémunération et les heures

L'Enquête sur l'emploi, la rémunération et les heures (EERH 1994) couvre tous les secteurs de l'industrie canadienne et recueille des données sur quatre grandes variables: i) les salaires et les paiements versés aux employés (notés z_2 ; ou paye); ii) le nombre d'employés (emploi); iii) le nombre d'heures travaillées par les employés (y_1 ; heures) et iv) le sommaire de la rémunération (y_2 ; rémunération). L'EERH recourt à un plan d'échantillonnage stratifié à deux phases. En un premier temps, on prélève un échantillon des comptes de retenues sur la paye au moyen d'un plan d'échantillonnage stratifié de Bernoulli, le taux d'échantillonnage des strates variant de 10 % à 100 %. Les strates sont établies par région, c'est-à-dire une ou plusieurs provinces. Nous décrivons l'estimation de l'EERH en fonction d'une région spécifique.

On commence par la transcription de deux variables, la paye (z_2) et le nombre d'employés (z_3), pour les unités

$$g_{1k} = N' / \hat{N}'_{1i} \quad (7.7)$$

pour toutes les valeurs $k \in s_{1i} = s_1 \cup U_i$, où $\hat{N}'_{1i} = \sum_{s_{1i}} w_{1k}$, $i = 1, \dots, I$.

Passons au deuxième calage. On utilise pour cela les groupes de calage s_{1ij} , $j = 1, \dots, J$, identifiés par le vecteur Δ_{2k} , qui apparaît en (5.3). Ces groupes reposent sur une classification par province et par industrie. Ils sont construits i) pour que y_k et les deux variables z -présentent une forte relation de régression et ii) pour que chaque groupe comprenne au moins 30 observations. Le vecteur x_k à $J(I+2)$ dimensions du calage de la deuxième phase est donné par

$$x'_k = \Delta'_{2k} \otimes (\Delta'_{1k}, z_{2k}, z_{3k}) \quad (7.8)$$

En vertu de cette spécification (voir le tableau 1), chaque valeur $k \in s_1$ peut être classée dans une des $I \times J$ cellules créées lors du croisement des groupes de calage de deux phases. Soit $s_{2j} = s_2 \cap s_{1j}$; $s_{1ij} = s_{1i} \cap U_i$; $s_{2ij} = s_2 \cap s_{1ij}$. On doit aussi connaître la valeur des variables quantitatives z_{2k} (paye) et z_{3k} (nombre d'employés) pour $k \in s_1$. Le vecteur spécifié x_k en (7.8) est complet parce qu'il intègre $x_{1k} = \Delta_{1k}$. Sa version réduite, qui exclurait les groupes de la première phase, correspondrait à $x'_k = \Delta'_{2k} \otimes (z_{2k}, z_{3k})$. Comme dans l'exemple 7.1, deux jeux de groupes de calage se recoupent.

Puisque le vecteur x_k de (7.8) a la même structure qu'en (5.8), on se sert de l'équation (5.9) pour dériver les facteurs g de la deuxième phase pour chaque groupe $j = 1, \dots, J$. De (7.8), il découle qu'on ajuste une régression distincte de y_k sur $z_k = (z_{2k}, z_{3k})'$ à chaque groupe de calage de la deuxième phase, les coordonnées à l'origine variant au niveau des groupes de calage de la première phase. Si $C_{2k} = 1$ pour toutes les valeurs k et si on retient la forme additive $g_k = g_{1k} + g_{2k} - 1$, pour l'ensemble des facteurs de calage, on obtient, après quelques transformations algébriques

$$g_k^* = G_{1i} G_{2ij} + H_j T_j^{-1} (z_k - z_{s_{2ij}})$$

pour toutes les valeurs $k \in s_{2ij}$, où

Remarque 6.2: Dans la pratique, on calcule rarement la variance estimée comme une double somme. Au terme de certaines transformations algébriques, il est possible de simplifier la double somme en somme simple dans certains plans d'échantillonnage importants. C'est le cas notamment de l'échantillonnage simple et de l'échantillonnage aléatoire simple stratifié, pour les deux phases. Sarnadal et coll. (1992) ont donné le développement algébrique explicite de la variance pour le premier cas, tandis que Hidiroglou (1995), et Binder, Babyak, Brodeur, Hidiroglou et Jocelyn (1997) en ont fait autant pour le second.

7. APPLICATIONS AVEC STRATIFICATION A POSTERIORI À LA PREMIÈRE PHASE

7.1 L'échantillon fiscal de Statistique Canada

Statistique Canada applique présentement l'approche du groupe de calage examinée à la partie 5 à son plan d'échantillonnage à deux phases des dossiers de l'impôt. L'exemple à son importance car il permet d'étendre la méthode classique de stratification a posteriori des plans d'échantillonnage à une phase aux plans d'échantillonnage à deux phases. Armstrong et Saint-Jean (1994) décrivent la technique d'échantillonnage, les critères de stratification a posteriori et les estimateurs. Nous verrons maintenant comment parvenir à ces derniers en prenant un cas particulier de la méthode présentée à la partie 5. Dans chaque phase, on utilise comme plan d'échantillonnage la stratification de Bernoulli, selon la technique du nombre aléatoire permanent. Chaque stratification s'appuie sur des critères différents. À cause de l'échantillonnage de Bernoulli, la taille de l'échantillon varie de façon aléatoire à chaque étape. Pour compenser la tendance à la hausse résultante de la variance, on procède à une stratification a posteriori aux deux phases de l'échantillonnage. Les deux critères de stratification diffèrent. En réalité, les deux stratifications se recoupent. Pour reprendre la terminologie de la partie 5, les strates a posteriori de la première phase correspondent aux groupes de calage de la première phase, notés $U'_i; i = 1, \dots, I$, et l'appartenance de l'unité k à un groupe est indiquée par le vecteur Δ_{1k} , donné en (5.1). Les strates a posteriori de la deuxième phase. On les note $s'_{ij}; j = 1, \dots, J$ et l'inclusion de l'unité k à un groupe est signalée par le vecteur Δ_{2k} qui apparaît en (5.3). On effectue le premier calage grâce à l'information sur la taille des strates a posteriori de la première phase N'_i . Cette enquête ne procure pas d'autres données complètement-aires donc $z_{1k} = 1$ pour toutes les valeurs k de (5.5), ce qui donne $x_{1k} = \Delta_{1k}$. En établissant que $C_{1k} = 1$ pour toutes les valeurs k , de (5.7) il découle que

$$g_{1k} = N'_i / N_{1i} \quad (7.1)$$

pour toutes les valeurs $k \in s'_{1i}$ où $N_{1i} = \sum_{k \in s'_{1i}} w_{1k}$ estime la taille connue de la strate a posteriori de la première phase N'_i

$$g_{k*} = \frac{N'_i}{N_i} \frac{N_{1i}}{N_{2i}} \quad (7.2)$$

pour toutes les valeurs de $k \in s_{2ij}$, où

$$N_{1i} = \sum_{i=1}^I \left(\frac{N_{1i}}{N_i} \right) N_{1ij}; N_{2j} = \sum_{i=1}^I \left(\frac{N_{1i}}{N_i} \right) N_{2ij} \quad (7.3)$$

avec $N_{1ij} = \sum_{k \in s_{1ij}} w_{1k}$ et $N_{2ij} = \sum_{k \in s_{2ij}} w_{2k}$. Dans le cas qui nous intéresse, $s_{2j} = s_{2i} \cap s_{1j}$ désigne la partie de l'échantillon s_2 de la deuxième phase qui tombe dans la strate a posteriori s_{1j} . Il s'ensuit que l'estimateur du total $Y(d)$ pour un domaine donné $U(d)$ correspond à $Y(d) = \sum_{k \in s_{2j}} g_{k*} Y_k(d)$, soit

$$Y(d) = \sum_{i=1}^I \sum_{j=1}^J \frac{N_{1i}}{N_i} \frac{N_{2j}}{N_{1i}} \sum_{k \in s_{2ij}} w_{2k} Y_k(d).$$

L'estimation de la variance exige deux résidus qu'on peut facilement tirer des expressions générales présentées à la partie 6.

Il existe d'autres solutions à la spécification du vecteur réduit $x_k = \Delta_{2k}$ et on s'en sert effectivement dans l'enquête. Voyons donc à quoi ressemble l'estimateur avec la spécification du vecteur complet. Pour le premier calage, supposons comme avant que $x_{1k} = \Delta_{1k}$ ce qui correspond à $z_{1k} = 1$ pour toutes les valeurs k de (5.8). L'équation (7.1) donne les facteurs g_{1k} de la première phase. Les données d'enquête disponibles permettent d'affecter chaque unité $k \in s_1$ à une des $I \times J$ cellules issues du croisement des deux critères de stratification a posteriori. Par conséquent, le vecteur x_k du deuxième calage peut s'exprimer de la manière suivante:

$$x'_k = \Delta_{1k} \otimes \Delta_{2k} \quad (7.4)$$

Il s'agit de la spécification du vecteur complet puisqu'elle inclut Δ_{1k} qui véhicule l'information de la première phase. Spécifions aussi que $C_{2k} = 1$ pour toutes les valeurs k . Puisque (7.4) emprunte la même forme que (5.8), les facteurs g_{2k} de la deuxième phase peuvent être tirés de (5.9) un groupe à la fois avec $z_k = \Delta_{1k}$. Les facteurs de calage globaux sont donnés par

$$g_{k*} = \frac{N'_i}{N_i} \frac{N_{1i}}{N_{2i}} \quad (7.5)$$

$$T_{2j} = \sum_{s_{2j}} \frac{C_{2k}}{w_{1k} w_{2k} z_k z_k'} \quad (5.10)$$

Les poids globaux $w_k^* = w_k^* g_k^*$ résultants où $g_k^* = g_{1k} g_{2k}$ sont les mêmes que ceux obtenus quand on procède au deuxième calage un groupe à la fois, le groupe j étant calé d'après la quantité $\sum_{s_{1j}} w_{1k} z_k$ qui est connue. Bref, $\sum_{s_{2j}} w_k^* z_k = \sum_{s_{1j}} w_{1k} z_k$ pour $j = 1, \dots, J$. Les groupes s_{1j} sont baptisés «groupes de calage de la deuxième phase». Nous sommes maintenant en mesure de calculer g_{1k} et g_{2k} groupe par groupe avec (5.7) et (5.9). On estime toujours le total Y avec (3.13).

6. ESTIMATION DES DOMAINES ET DE LA VARIANCE

Les parties qui précèdent traitaient de l'estimation du total de y au niveau de la population prise dans son ensemble. La plupart des enquêtes nécessitent aussi cependant des estimations pour divers domaines ou sous-populations auxquels on s'intéresse. Les demandes concernant l'estimation d'un domaine peuvent surgir avant ou après l'échantillonnage. Les données auxiliaires jouent un rôle essentiel pour les domaines. On pourra se procurer l'estimation précise d'un domaine (même petit) i) si les groupes de calage et les domaines d'intérêt concordent étroitement, et ii) si les variables auxiliaires illustrent un fort lien de régression avec la ou les variables qui suscitent l'intérêt.

Notons $U_d^a (U_d \subset U)$ un domaine de la population U qu'on aimerait estimer. Le total y du domaine U_d se définit par $Y(d) = \sum_{U_d} y_k = \sum_{U_d} y_k(d)$ où $y_k(d) = y_k$ si $k \in U_d$ et $y_k(d) = 0$ si $k \notin U_d$.

L'estimateur de $Y(d)$ est

$$\hat{Y}(d) = \sum_{s_2} w_k^* y_k(d) \quad (6.1)$$

où les poids globaux calés $w_k^* = w_k^* g_k^*$ peuvent être calculés un groupe à la fois tel qu'indiqué à la partie 5. On établit les facteurs de calage g_{1k} et g_{2k} à partir de toutes les données auxiliaires disponibles, ainsi qu'on le mentionne au tableau 1. Dans ce sens, les poids globaux calés w_k^* qui en résultent sont donc les meilleurs qui soient. Remarquons que ces poids sont indépendants des domaines particuliers de l'enquête à estimer.

Pour établir l'estimateur de la variance de l'estimateur total du domaine $\hat{Y}(d)$, on recourt à une méthode reposant sur le plan d'échantillonnage. En d'autres mots, la variance est interprétée en fonction d'un tirage répétitif d'échantillons s_1 et s_2 . Särndal et coll. (1992) donnent des précisions sur cette technique de dérivation (résultat 9.7.1 p. 362). Les probabilités d'inclusion du premier et du deuxième ordre sont intégrées aux poids qui entrent dans la formule de la variance. Les poids associés à l'échantillon de la première phase sont $w_{1k} = 1/\pi_{1k}$ et $w_{1k\ell} = 1/\pi_{1k\ell}$ où $\pi_{1k\ell} = P(k \text{ and } \ell \in s_1)$. Ils ont pour contrepartie $w_{2k} = 1/\pi_{2k}$

$$Y(d) \text{ correspond à } \sum_{k \in s_2} \sum_{k \in s_2} w_{2k} (w_{1k} w_{1\ell} - w_{1k\ell}) (g_{1k} e_{1k}(d)) (g_{1\ell} e_{1\ell}(d)) - \sum_{k \in s_2} \sum_{k \in s_2} w_{1k} w_{1\ell} (w_{2k} w_{2\ell} - w_{2k\ell}) (g_{2k} e_{2k}(d)) (g_{2\ell} e_{2\ell}(d)) \quad (6.2)$$

Souignons que lorsque $k = \ell$ nous avons $w_{1k\ell} = w_{1k}$ et $w_{2k\ell} = w_{2k}$ en (6.2). Spécifions maintenant les résidus de régression de (6.2), en présupmant qu'il existe des groupes de calage de première phase $U_i, i = 1, \dots, I$, et de deuxième phase $s_{1j}, j = 1, \dots, J$, ainsi qu'on l'a expliqué à la partie 5. Les sous-ensembles associés à l'échantillon sont dénotés ainsi: $s_{2i} = s_2 \cap U_i, s_{2j} = s_2 \cap s_{1j}$. Les valeurs résiduelles nécessaires à (6.2) sont les suivantes, pour $k \in (s_{2i} \cap U_d)$,

$$e_{1k}(d) = y_k(d) - z_{1k}' \hat{b}_{1i}(d) \quad (6.3)$$

et pour $k \in (s_{2j} \cap U_d)$

$$e_{2k}(d) = y_k(d) - z_{2k}' \hat{b}_{2j}(d) \quad (6.4)$$

Les vecteurs de régression estimés $\hat{b}_{1i}(d)$ et $\hat{b}_{2j}(d)$ correspondent à

$$\hat{b}_{1i}(d) = T_{1i}^{-1} \left\{ \sum_{s_{1i}} \frac{C_{1k}}{w_{1k} z_{1k} y_{2k}(d)} + \sum_{s_{2j}} \frac{C_{1k}}{w_k^* z_{1k} (y_k(d) - y_{2k}(d))} \right\} \quad (6.5)$$

où T_{1i} vient de (5.6), et

$$\hat{b}_{2j}(d) = T_{2j}^{-1} \sum_{s_{2j}} \frac{C_{2k}}{w_{1k} w_{2k} z_k y_k(d)} \quad (6.6)$$

avec T_{2j} est donné par (5.10), et

$$y_{2k}(d) = z_k' \hat{b}_{2j}(d) \text{ pour } k \in (s_{2j} \cap U_d).$$

Remarque 6.1: Notons que pour chaque nouveau domaine d'intérêt, l'estimateur de la variance (6.2) nécessite deux nouveaux jeux de résidus dépendants $e_{1k}(d)$ et $e_{2k}(d)$. On en a également besoin pour toutes les unités k de l'échantillon s_2 de la deuxième phase, y compris les unités n'appartenant pas au domaine. Estimer la variance d'un domaine peut donc s'avérer fastidieux.

pas identique à (3.13) et s'avère moins efficace, car $B_{1,alt}$ exploite moins d'information sur x_{1k} que B_1 .

5. GROUPES DE CALAGE

Nous appliquerons ici les résultats des parties 3 et 4 au cas important où les données auxiliaires du tableau 1 renferment des précisions sur des sous-ensembles complets de la population U et de l'échantillon de la première phase s_1 s'excluant mutuellement. Ces sous-ensembles sont désignés U_i , $i = 1, \dots, I$, et ceux de l'échantillon de la première phase, s_{1j} , $j = 1, \dots, J$. On les appelle des groupes de calage, pour des raisons qui deviendront plus claires par la suite. Les strates a posteriori sont une forme simple de groupe de calage.

Les vecteurs Δ_{1k} et Δ_{2k} indiqueront l'appartenance de l'unité k aux groupes de calage U_i et s_{1j} , respectivement. Ces groupes sont identifiés par

$$\Delta_{1k} = (\delta_{11k}, \dots, \delta_{1Ik})' \quad (5.1)$$

avec

$$\delta_{1ik} = \begin{cases} 1 & \text{si } k \in U_i \\ 0 & \text{autrement} \end{cases} \quad \text{pour } i = 1, \dots, I \quad (5.2)$$

et

$$\Delta_{2k} = (\delta_{21k}, \dots, \delta_{2Jk})' \quad (5.3)$$

avec

$$\delta_{2jk} = \begin{cases} 1 & \text{si } k \in s_{1j} \\ 0 & \text{autrement} \end{cases} \quad \text{pour } j = 1, \dots, J \quad (5.4)$$

Outre l'information concernant l'inclusion au groupe, qui est qualitative et établie par Δ_{1k} et Δ_{2k} , on pourrait disposer d'informations sur les variables quantitatives (continues ou discrètes) de l'unité k . Nous les appelons «variables auxiliaires complémentaires». Les données discrètes sur une unité (entreprise) venant d'une enquête auprès des entreprises pourraient comprendre un code d'industrie ou un code d'emplacement. L'information sur les variables quantitatives pourrait aussi se rapporter au nombre d'employés ou au chiffre d'affaires brut de l'unité. Certaines variables auxiliaires complémentaires pourraient être connues pour l'ensemble de la population ou seulement pour l'échantillon de la première phase.

Nous supposons ici que le vecteur x_{1k} , servant à établir les facteurs g de la première phase est structuré comme suit

$$x_{1k}' = \Delta_{1k}' \otimes z_{1k}' \quad (5.5)$$

où z_{1k} de la dimension Q_1 correspond au vecteur des variables auxiliaires complémentaires dont on dispose pour

l'échantillon de la première phase. Les exigences en matière d'information du tableau 1 s'appliquent au vecteur x_{1k} . En d'autres mots, il faut savoir soit l'appartenance au groupe indiquée par Δ_{1k} et la valeur de z_{1k} pour tous les éléments $k \in U$, soit le total $\sum_{U_i} z_{1k}$ de chaque groupe pris séparément, $i = 1, \dots, I$.

Quand x_{1k} adopte la forme présentée en (5.5), on peut établir les facteurs g_{1k} de (3.5) du groupe, un groupe à la fois. La matrice T_1 de (3.6) à inverser est une matrice quasi-diagonale à IQ_1 dimensions. La matrice quasi-diagonale type de dimensions Q_1 par Q_1 est représentée par

$$T_{1i} = \sum_{s_{1i}} \frac{C_{1k}}{w_{1k} z_{1k} z_{1k}'} \quad (5.6)$$

pour $i = 1, \dots, I$. L'inverse de T_1 est aussi une matrice quasi-diagonale avec pour matrices diagonales $(T_{1i})^{-1}$. Les blocs de la matrice inverse de T_1 qui ne se trouvent pas sur la diagonale sont des matrices nulles. De (3.6), on tire donc

$$g_{1k} = 1 + \left(\sum_{U_i} z_{1k} - \sum_{s_{1i}} w_{1k} z_{1k} \right) (T_{1i})^{-1} \frac{C_{1k}}{z_{1k}} \quad (5.7)$$

pour $k \in s_{1i}$, $i = 1, \dots, I$, où T_{1i} vient de (5.6). Notons que les poids w_{1k} résultants sont identiques à ceux obtenus quand on procède au premier calage un groupe à la fois, le groupe i étant calé sur le total connu $\sum_{U_i} z_{1k}$, c'est-à-dire $\sum_{s_{1i}} w_{1k} z_{1k} = \sum_{U_i} z_{1k}$ pour $i = 1, \dots, I$. L'appellation «groupes de calage de la première phase» convient donc parfaitement aux groupes U_i .

Examinons maintenant les facteurs g_{2k} de la deuxième phase donnés en (3.11). Ces facteurs reposent sur les vecteurs auxiliaires x_k , qu'on doit connaître pour les unités $k \in s_1$. On présume que x_k renferme de l'information sur les groupes de la deuxième phase, de sorte que

$$x_k' = \Delta_{2k}' \otimes z_k' \quad (5.8)$$

où Δ_{2k} représente l'identificateur des groupes de la deuxième phase et z_k est la valeur du vecteur des variables auxiliaires complémentaires disponibles, pour $k \in s_1$. Puisque les exigences du tableau 1 s'appliquent, Δ_{2k} (l'appartenance aux groupes de la deuxième phase) et la valeur de z_k (le vecteur auxiliaire complémentaire) doivent être connus pour toutes les valeurs $k \in s_1$. Dans ce cas, z_k peut comprendre une partie ou la totalité de l'information de x_{1k} fournie par (5.5), ainsi que tout autre renseignement disponible pour les unités $k \in s_1$.

Quand x_k a la structure indiquée en (5.8), on peut aussi obtenir les facteurs g_{2k} en les calculant groupe par groupe. Le fait que la matrice à inverser dans (3.11) soit une matrice quasi-diagonale autorise pareille simplification. On parvient donc à

$$g_{2k} = 1 + \left(\sum_{s_{1j}} w_{1k} z_k - \sum_{s_{2j}} w_{1k} w_{2k} z_k \right) (T_{2j})^{-1} \frac{C_{2k}}{z_k} \quad (5.9)$$

pour $k \in s_{2j} = s_2 \cap s_{1j}$, $j = 1, \dots, J$, où

toutefois pas calés de s_2 à s_1 , parce que x_{1k} n'intervient pas dans le deuxième calage. Par conséquent, $\sum_{s_2}^k \tilde{w}_{1k}^* x_{1k} \neq \sum_{s_1}^k \tilde{w}_{1k}^* x_{1k} = \sum_U x_{1k}$. Si l'enquête requiert un système de pondération reproduisant la somme $\sum_U x_{1k}$ connue, on doit donc recourir à la spécification du vecteur complet.

Jusqu'à présent, nous sommes restreints au cadre général du calage quand existent deux couches de données auxiliaires. Ce cadre ne révèle toutefois rien des nombreuses formes intéressantes que peut prendre l'estimateur \hat{Y} de (3.13) avec des données d'un certain type. La partie 7 fournit quelques illustrations. Nous commencerons par aborder trois points d'intérêt pratique pour virtuellement n'importe quelle grande enquête: i) la stratification a posteriori ou, d'une manière plus générale, l'existence d'information auxiliaire sur divers sous-groupes de la population (partie 5), ii) l'estimation de domaines d'intérêt (partie 6) et iii) l'estimation de la variance (partie 6).

4. L'ESTIMATEUR DE CALAGE À DEUX PHASES EN TANT QU'ESTIMATEUR DE RÉGRESSION

L'estimateur de calage (3.13) peut être exprimé autrement grâce à la formule (4.1) qui suit. La nouvelle expression lie l'estimateur exactement à l'estimateur de régression pour les plans d'échantillonnage à deux phases présentés par Särndal et coll. (1992, chapitre 9).

Théorème 4.1: Quand les poids globaux \tilde{w}_k^* calés sont établis au moyen de l'équation (3.9), l'estimateur de calage (3.13) est identique à l'estimateur de régression à deux phases

$$\hat{Y} = \sum_U y_{1k} + \sum_{s_1} w_{1k} (y_{2k} - y_{1k}) + w_k^* (y_k - y_{2k}) \quad (4.1)$$

où y_{1k} et y_{2k} correspondent aux prévisions de régressions successives, de sorte que

$$y_{1k} = x_{1k}' \hat{B}_1 \quad (4.2)$$

avec

$$\hat{B}_1 = T_1^{-1} \left\{ \sum_{s_1} \frac{C_{1k}}{w_{1k} x_{1k}' y_{2k}} + \sum_{s_2} \frac{C_{1k}}{w_k^* x_{1k} (y_k - y_{2k})} \right\} \quad (4.3)$$

et T_1 est donné par (3.6), tandis que

$$y_{2k} = x_k' \hat{B}_2 \quad (4.4)$$

où

$$\hat{B}_2 = T_2^{-1} \sum_{s_2} \frac{C_{2k}}{\tilde{w}_{1k} w_{2k} x_k y_k} \quad (4.5)$$

et T_2 est donné par (3.12).

La preuve de ce qui précède nécessite des calculs algébriques fastidieux mais simples. Nous ne la reproduisons donc pas ici.

Nous verrons maintenant qu'on peut arriver à (4.1) en deux étapes, au moyen de l'estimation par régression. En premier lieu, supposons que la variable y_k à laquelle on s'intéresse a été observée pour l'ensemble de l'échantillon de la première phase s_1 . On dispose de données auxiliaires sur x_{1k} pour $k \in s_1$ et connaît le total de la population $\sum_U x_{1k}$. L'estimateur de régression de $y = \sum_U y_k$ serait donc

$$\hat{Y} = \sum_U y_{1k}^0 + \sum_{s_1} w_{1k} (y_k - y_{1k}^0) \quad (4.6)$$

Le premier terme de la dernière expression correspond à l'estimateur Horvitz-Thompson (hypothétique) de Y pour la première phase. Le deuxième et le troisième termes représentent un ajustement de régression en vertu duquel on prédit y_{1k}^0 selon la régression de y_k ajustée sur x_{1k} , pour $k \in s_1$. Bref, $y_{1k}^0 = x_{1k}' \hat{B}_0^1$, où

$$\hat{B}_0^1 = T_1^{-1} \sum_{s_1} \frac{C_{1k}}{w_{1k} x_{1k} y_k}.$$

Souignons que $\sum_U y_{1k}^0 = (\sum_U x_{1k})' \hat{B}_0^1$ où $\sum_U x_{1k}$ est connu. Aucun des termes de (4.6) ne peut toutefois être calculé directement car on n'observe y_k que dans l'échantillon de la deuxième phase. Il faut donc procéder à une autre estimation par régression. Pour cela, on remplace le terme $\sum_{s_1} w_{1k} y_k$ inconnu de (4.6) par l'estimateur de régression conditionnel

$$\sum_{s_1} w_{1k} y_{2k}^* + \sum_{s_2} w_k^* (y_k - y_{2k}^*) \quad (4.7)$$

où $y_{2k}^* = x_k' \hat{B}_2$ et \hat{B}_2 donné dans (4.5), prédit y_k d'après la régression de y_k sur x_k , qu'on connaît jusqu'à s_1 . Le vecteur \hat{B}_0^1 dont on a besoin pour calculer y_{1k}^0 inclut une matrice T_1 connue et un vecteur inconnu

$$\sum_{s_1} \frac{C_{1k}}{w_{1k} x_{1k} y_k}.$$

En appliquant un estimateur de régression au vecteur inconnu, on obtient \hat{B}_1 , qui apparaît dans (4.3) et peut remplacer \hat{B}_0^1 . Ces deux substitutions dans (4.6) aboutissent à l'estimateur de régression à deux phases de (4.1), qui est identique à l'estimateur de calage (3.13).

Remarque 4.1: Une solution plus directe à \hat{B}_1 en (4.3) consisterait à n'utiliser que l'échantillon de la deuxième phase. On aurait ainsi obtenu

$$\hat{B}_{1,alt} = \left(\sum_{s_2} \frac{C_{2k}}{w_k^* x_{1k} x_{1k}'} \right)^{-1} \sum_{s_2} \frac{C_{2k}}{w_k^* x_{1k} y_k}.$$

Les prévisions résultantes $y_{1k,alt} = x_{1k}' \hat{B}_{1,alt}$ remplaceraient y_{1k} dans (4.1). L'estimateur de régression n'est néanmoins

sous réserve de l'équation de calage de la deuxième phase

$$(3.8) \quad \sum_{s_2} w_k^* x_k = \sum_{s_1} w_{1k} x_k$$

où $x_k = (x'_{1k}, x'_{2k})'$. Les poids globaux calés qui en résultent sont

$$(3.9) \quad w_k^* = w_k^* g_k^*$$

où

$$(3.10) \quad g_k^* = g_{1k} g_{2k}$$

et g_{1k} est issu de (3.5), tandis que g_{2k} correspond à

$$(3.11) \quad g_{2k} = 1 + \left(\sum_{s_1} w_{1k} x_k - \sum_{s_2} w_{1k} w_{2k} x_k \right) \left(T_2^{-1} \frac{C_{2k}}{x_k} \right)$$

pour $k \in s_2$, et

$$(3.12) \quad T_2 = \sum_{s_2} \frac{w_{1k} w_{2k} x_k x'_k}{C_{2k}}$$

Une fois de plus, il arrive que g_k^* ait une valeur nulle ou négative, mais on peut faire en sorte qu'elle soit positive en ajoutant à (3.8) l'inégalité $w_k^* > 0$ pour $k \in s_k$. Après avoir déterminé les poids globaux w_k^* grâce à l'équation (3.9), on obtient l'estimateur de Y

$$(3.13) \quad \hat{Y} = \sum_{s_2} w_k^* y_k$$

Remarque 3.1: L'approche qui précède peut soulever un problème. En effet, il se peut que g_{1k} prenne une valeur nulle ou négative, auquel cas (3.7) ne permet pas de mesurer l'écart. Ce problème ne se pose pas pour quelques grandes applications comme la stratification a posteriori, car les valeurs g_{1k} qui s'y associent sont toujours supérieures à zéro. Si g_{1k} dépasse toujours zéro, on peut accepter le critère de minimisation établi par (3.7). Sinon, il doit être modifié. Une solution éventuelle consiste à imposer les contraintes précitées pour que w_{1k} ait une valeur positive pour les unités $k \in s_1$. Une autre serait de remplacer C_{2k} dans (3.7) par

$$C_{2k}^* = C_{2k} \frac{w_{1k}}{w_{1k}^*}$$

Il s'ensuivrait que

$$\frac{C_{2k}^*}{w_{1k}^* w_{2k}^*} = \frac{C_{2k}}{w_k^*},$$

qui est toujours positif. On peut montrer que les facteurs g_k^* de (3.9) correspondent à $g_k^* = g_{1k} + g_{2k} - 1$, où g_{1k} venant de (3.5), comme avant, et g_{2k} de (3.11), pourvu qu'on définisse plutôt ainsi T_2 comme

$$T_2 = \sum_{s_2} \frac{w_k^* x_k x'_k}{C_{2k}}.$$

À notre avis, qu'on choisisse la forme multiplicative $g_k^* = g_{1k} g_{2k}$ ou la forme additive $g_k^* = g_{1k} + g_{2k} - 1$ les estimations ne s'en ressentiront guère dans la majorité des applications. Nous pensons en effet que les deux estimations ponctuelles seront voisines et qu'il en ira autant pour l'estimation de la variance.

Remarque 3.2: Bornez les poids n'a habituellement qu'une incidence négligeable sur les estimations. Comme on peut le lire dans Stukel, Hidiroglou et Särndal (1996), les travaux récents sur le calage dans le cadre de plans d'échantillonnage à une phase montrent que des jeux de poids g légèrement distincts débouchent sur les estimations ponctuelles presque identiques. Certains ont dernièrement mis au point des logiciels de calage, notamment celui décrit par Deville et coll. (1993), qui minimise une fonction de distance afin que les facteurs g résultants soient bornés en haut et en bas.

Remarque 3.3: Les données auxiliaires du tableau 1 peuvent être utilisées de plusieurs manières pour le calage à deux phases. Si on prend notamment l'équation de calage de la deuxième phase définie en (3.8), trois spécifications possibles pour le vecteur x_k sont: i) $x_k = (x'_{1k}, x'_{2k})'$; ii) $x_k = x'_{2k}$; et iii) $x_k = x'_{1k}$. Voici ce que nous pensons de ces possibilités, qui donnent les poids calés de la première phase de (3.4) après calage à la première phase.

La spécification i) $x_k = (x'_{1k}, x'_{2k})'$, que préconisent Särndal et coll. (1992), exploite toute l'information disponible. À cet égard, la spécification est donc idéale. Les cas ii) et iii) négligent une partie des données. Le cas ii) présente parfois de l'intérêt, même si certaines données sont perdues; on en trouve un exemple à la partie 7.1. Le cas iii) implique que les données $\{x_{2k} : k \in s_1\}$ sont observées mais qu'on ne s'en sert pas. Nous ne nous y attarderons donc pas davantage. Nous avons baptisé $x_k = (x'_{1k}, x'_{2k})'$ le vecteur complet et $x_k = x'_{2k}$ le vecteur réduit.

On peut procéder au calage de la deuxième phase du vecteur réduit $x_k = x'_{2k}$ sans perdre beaucoup d'information si x'_{2k} constitue un bon substitut à x'_{1k} , ainsi que l'a aussi observé Dupont (1995). Si x'_{1k} complète x'_{2k} , cependant, on devrait plutôt se servir du vecteur complet $x_k = (x'_{1k}, x'_{2k})'$ pour le calage décrit en (3.7). Sans cela, on perdra une partie appréciable de l'information et la variance pourrait augmenter.

Remarque 3.4: Les vecteurs complets et réduits x_k donnent les poids globaux w_k^* calés sur x'_{2k} , de s_2 à s_1 . En d'autres termes, $\sum_{s_1} w_k^* x'_{2k} = \sum_{s_1} w_{1k} x'_{2k}$, car (3.8) se vérifie et x'_{2k} est inclus dans x_k . Le vecteur complet et le vecteur réduit se distinguent néanmoins pour ce qui est du calage fondé sur x'_{1k} . Si on se sert de la spécification du vecteur complet à la deuxième étape, les poids globaux sont calés d'après x'_{1k} de s_2 à s_1 et de s_1 à U . Bref, $\sum_{s_2} w_k^* x'_{1k} = \sum_{s_1} w_{1k} x'_{1k} = \sum_U x'_{1k}$. Avec la spécification du vecteur réduit, en revanche, les poids globaux w_k^* résultants sont calés sur x'_{1k} de s_1 à U à cause du calage de la première phase. En d'autres termes, $\sum_{s_1} w_{1k} x'_{1k} = \sum_U x'_{1k}$. Ils ne sont

le font Srndal et coll. (1992, chapitre 9), divisons x_k en $x_k = (x'_{1k}, x'_{2k})'$. L'information sur le vecteur x'_{1k} touche l'ensemble de la population, mais dans le cas du vecteur x'_{2k} , elle ne concerne que l'chantillon de la premire phase. Le tableau 1 rsume les hypothses qui prcdent sur les donnes auxiliaires  notre disposition pour l'estimation.

Tableau 1

Liens entre les jeux d'units et donnes disponibles aux diffrents niveaux

Jeu d'units	Donnes existantes
Population	$\{x_{1k} : k \in U\}$ or $\sum_U x_{1k}$
chantillon de la premire phase	$\{x_k : k \in s_1\}$
chantillon de la deuxime phase	$\{(x_k, y_k) : k \in s_2\}$

Remarquons qu'on n'a pas besoin des valeurs individuelles x_{1k} , $k \in U$. Il suffit de conntre le total $\sum_U x_{1k}$, qu'on peut extraire d'une source administrative fiable. L'existence de donnes auxiliaires  une phase quelconque ou aux deux ouvre la porte  une modification des poids d'chantillon-nage  l'aide de facteurs de calage calculs d'aprs l'information complmentaire. On modifie les poids d'chantillon-nage d'une unit  chaque phase en le multipliant par le facteur de calage, ce qui donne le poids calage. Le poids de calage w_{1k} de la premire phase pour les units $k \in s_1$ correspond  $w_{1k} = w_{1k} g_{1k}$. Le poids d'chantillon-nage de la premire phase est w_{1k} , et le facteur de calage correspondant, g_{1k} . De mme, on calcule les poids de calage globaux $w_k^* = w_k g_k^*$ des units $k \in s_2$, o g_k^* reprsente le facteur de calage global. L'exposant «*» dsigne les poids globaux intgrant les deux phases. Le symbole «~» qui s'y rajoute prcise que les poids sont cals.

3. CALAGE PAR LA DISTANCE GNRALISE DES MOINDRES CARRS

Les donnes auxiliaires disponibles  chaque tape de l'chantillon-nage peuvent servir  amliorer les poids  l'aide d'un processus de calage. La variance des estimations rsultantes s'en trouvera rduite si les variables auxiliaires et celles auxquelles on s'intresse sont troitement corrles. Nous dsirons un autre jeu de poids qui se rapprochera au maximum des poids initiaux. Pour procder au calage, on doit spcifier une mesure de l'cart entre les poids initiaux et les nouveaux poids. Plusieurs fonctions de distance ont t avances; lire  ce sujet Deville et Srndal (1992), Deville, Srndal, et Sautory (1993), et Singh et Mohi (1996). Toutes ces fonctions pourraient servir au calage  deux phases cependant, nous nous limiterons  une seule d'entre elles, soit la fonction gnralise des moindres carrs (GMC). Pour un ensemble arbitraire d'units s , cette fonction prend la forme

$$D = \frac{1}{2} \sum_s C_k \frac{w_k}{(w_k - w_k^*)^2} \tag{3.1}$$

i) **Premier calage** (de s_1  U). Les poids d'chantillon-nage de la premire phase $\{w_{1k} : k \in s_1\}$ servent de poids initiaux. Soit $\{C_{1k} : k \in s_1\}$, les facteurs positifs prtablis. On dtermine les poids cals de la premire phase en minimisant la distance GMC

$$D_1 = \frac{1}{2} \sum_{s_1} C_{1k} \frac{w_{1k}}{(w_{1k} - w_{1k}^*)^2} \tag{3.2}$$

sous rserve de l'quation de calage de la premire phase

$$\sum_{s_1} w_{1k} x_{1k} = \sum_U x_{1k} \tag{3.3}$$

o le total $\sum_U x_{1k}$ est connu. Notons que ce calage ne fait pas intervenir l'information sur x_{2k} car elle n'est disponible que pour s_1 . Les poids rsultants sont

$$w_{1k} = w_{1k} g_{1k} \tag{3.4}$$

avec

$$g_{1k} = 1 + \left(\sum_U x_{1k} - \sum_{s_1} w_{1k} x_{1k} \right) T_1 \frac{C_{1k}}{x_{1k}} \tag{3.5}$$

et

$$T_1 = \sum_{s_1} \frac{C_{1k}}{w_{1k} x_{1k} x'_{1k}} \tag{3.6}$$

Une partie des valeurs w_{1k} obtenues en (3.4) peuvent tre ngatives ou nulles. Maints utilisateurs prfrent travailler avec des poids positifs. On peut rectifier la situation en greffant l'ingalit $w_{1k} > 0$  (3.3) pour toutes les valeurs $k \in s_1$. Contrairement  ceux de (3.4), les poids rsultants ne s'expriment pas sous une forme explicite.

ii) **Deuxime calage** (de s_2  s_1).

On se sert de $\{w_{1k}, k \in s_2\}$ comme poids initiaux, w_{1k} venant de (3.4). Ces poids intgrent l'information sur x_{1k} pour l'ensemble de la population. En les appliquant aux donnes $\{y_k : k \in s_2\}$, on obtient un estimateur envisageable, soit $\bar{Y} = \sum_{s_2} w_{1k} w_{2k} y_k$. Nanmoins, puisqu'ils n'intgrent pas l'information existante sur la valeur de x_{2k} dont on dispose pour $k \in s_1$, ces poids peuvent tre amliors par un deuxime calage. Soit $\{C_{2k} : k \in s_2\}$, les facteurs positifs spcifiques. On obtient les poids globaux cals w_k^* en minimisant

$$D_2 = \frac{1}{2} \sum_{s_2} C_{2k} \frac{w_{1k} w_{2k}}{(w_{1k}^* - w_{1k} w_{2k})^2} \tag{3.7}$$

calage peut être exprimé sous la forme d'un estimateur de régression à deux phases parfaitement équivalent, c'est-à-dire d'un estimateur issu de deux ajustements par régression successifs. D'autres résultats théoriques apparaissent aux parties 5 et 6. La première traite des formes de l'estimateur de calage à deux phases pour des données importantes d'un type particulier, en l'occurrence quand certaines variables auxiliaires, que ce soit à la première ou à la deuxième étape, correspondent aux variables nominales codifiant un groupe d'éléments en catégories complètes qui s'excluent mutuellement. La partie 6 fournit des précisions sur deux aspects qui suscitent toujours beaucoup d'intérêt dans les enquêtes et occupent une place prépondérante dans le SGF, soit a) l'estimation des domaines (sous-populations) et b) l'estimation de la variance attribuable au plan d'échantillonnage. Pour estimer la variance, nous avons retenu l'approche de Särndal et Swensson (1987). À la partie 7, on verra comment la théorie exposée antérieurement trouve application à Statistique Canada. Finalement, la partie 8 récapitule brièvement ce qui a été appris.

2. NOTATION

La population est représentée par $U = \{1, \dots, k, \dots, N\}$. On prélève un premier échantillon probabiliste $s_1 (s_1 \subseteq U)$ selon un plan d'échantillonnage pour lequel la probabilité de sélection est $\pi_{1k} = P(k \in s_1 | s_1)$. Notons qu'il s'agit de probabilités conditionnelles, puisqu'elles supposent qu'on connaît s_1 . On présume que $\pi_{1k} > 0$ pour toutes les valeurs $k \in U$ et que $\pi_{2k} > 0$ pour toutes les valeurs $k \in s_1$. À partir de là, nous nous servirons des poids pour l'estimation. Le poids de l'unité k sera noté $w_{1k} = 1/\pi_{1k}$ pour l'échantillon de la première phase et $w_{2k} = 1/\pi_{2k}$ pour celui de la deuxième phase. Le poids d'échantillonnage global d'une unité quelconque sera donc $w_k^* = w_{1k} w_{2k}$.

L'objectif consiste à estimer le total de population $Y = \sum_{U} y_k$, où y_k représente la valeur de la variable y à laquelle on s'intéresse pour l'unité k . Si $A \subseteq U$ désigne un lieu de $\sum_{k \in A}^k$. Habituellement, l'échantillonnage à deux phases exige la collecte de données peu coûteuses sur les unités k qui appartiennent au vaste échantillon s_1 de la première phase. On se sert ensuite de l'information recueillie pour procéder à un échantillonnage et à une estimation très efficaces, à la deuxième étape. Les valeurs de y_k sont saisies pour $k \in s_2$. La formule $\bar{Y} = \sum_{s_2} w_k^* y_k$ procure un estimateur non biaisé de Y , reposant uniquement sur les poids d'échantillonnage. Nous examinerons maintenant les estimateurs de régression, qui autorisent une meilleure exploitation des données auxiliaires existantes. Appelons x le vecteur auxiliaire de l'échantillon de la première phase et sa valeur pour l'unité k , en x_k . Comme

nouveaux poids, dits «calés». Le poids calé d'un élément correspond au produit du poids initial par un facteur de calage qu'on obtient en minimisant une fonction qui mesure l'écart entre les poids initiaux et les poids calés, sous réserve que ces derniers fournissent une estimation exacte des totaux auxiliaires connus de la population. Avec l'échantillonnage à deux phases, l'existence de deux couches d'information signifie deux calages consécutifs. Le premier repose sur les données auxiliaires disponibles (au moins les chiffres de la population) et s'effectue au niveau de la population, dans son ensemble, ce qui donne les poids calés de la première phase. Le deuxième calage s'appuie sur les poids calés de la première phase en y intégrant les données sur l'échantillon de la première phase. Il en résulte une série de poids calés finaux.

Les deux méthodes tirent parti des deux couches d'information, mais n'aboutissent pas nécessairement à des résultats identiques. On le doit à la façon dont les ajustements par régression et l'approche de calage sont formulés. Dupont (1995) l'illustre clairement. En effet, elle a mis au point quatre estimateurs par l'approche de la régression. Ces estimateurs diffèrent en ce sens que les variables auxiliaires servent à dériver la valeur prévue de y nécessaire à l'estimateur de régression. Dupont établit l'estimateur correspondant à chacune des quatre variantes par calage, mais n'atteint une parité équivalente entre les deux approches que dans un cas sur quatre. Trois des quatre variantes examinées par Dupont peuvent être considérées comme des cas particuliers de l'approche générale exposée ici.

Partant des travaux de Hidiroglou et de Särndal (1995), nous proposons une théorie intégrée pour l'échantillonnage à deux phases en présence de données auxiliaires. Nous montrerons qu'il est possible d'obtenir les estimateurs de régression parce qu'ils forment un cas particulier de la méthode de calage. Les deux méthodes présentent donc un lien direct. Nous nous sommes intéressés à ce travail en vue de fournir les outils nécessaires à une exploitation efficace des sources de données administratives dans le cadre de plusieurs grandes enquêtes de Statistique Canada. Ce travail pave aussi la voie à l'inclusion de la technique d'échantillonnage à deux phases au Système généralisé d'estimation (SGE) de Statistique Canada, décrit dans Estevao, Hidiroglou et Särndal (1995).

Nous illustrerons notre théorie générale en l'appliquant à deux plans d'échantillonnage actuellement en usage à Statistique Canada. Armstrong et Saint-Jean (1994) parlent de la première application, soit l'échantillonnage à deux phases des dossiers de l'impôt. La seconde, mentionnée dans Hidiroglou, Latouche, Armstrong et Gossen (1995), concerne l'échantillonnage à deux phases des comptes de retenues sur la paye dans le cadre de l'Enquête sur l'emploi, la rémunération et les heures de Statistique Canada. L'article se structure comme suit. La partie 2 expose la notation. La partie 3 spécifie la version de la méthode de calage appliquée à l'échantillonnage à deux phases. À la partie 4 est présentée le résultat capital que l'estimateur de

Emploi des données auxiliaires dans l'échantillonnage à deux phases

M.A. HIDIROGLOU et C.-E. SÄRNDAL¹

RÉSUMÉ

Les plans d'échantillonnage à deux phases permettent d'utiliser les données auxiliaires de diverses façons. Les auteurs débütent en passant en revue les différents aspects que peuvent prendre ces données dans les enquêtes à deux phases. Ils établissent ensuite la méthode en vertu de laquelle elles sont converties en poids calés dont on se sert pour créer de bons estimateurs d'un total d'une population. Le calage s'effectue en deux étapes: i) au niveau de la population et ii) à celui de l'échantillon de la première phase. Les auteurs montrent qu'on peut aussi dériver les estimateurs issus de calage par régression, également en deux temps. Ils examinent ces estimateurs dans un cas particulier, en l'occurrence quand les données auxiliaires portent sur quelques sous-ensembles de la population baptisés «groupes de calage». Les strates a posteriori en constituent l'illustration la plus simple. Suit une discussion sur l'estimation des domaines d'intérêt et de la variance. Enfin, les résultats sont appliqués à deux importants plans d'échantillonnage à deux phases en usage à Statistique Canada. La théorie générale concernant l'emploi des données auxiliaires dans l'échantillonnage à deux phases sera intégrée au Système généralisé d'estimation de Statistique Canada.

MOTS CLÉS: Régression généralisée; échantillonnage à deux phases; approche assistée par modèle; estimation des domaines; facteurs de calage.

1. INTRODUCTION

L'échantillonnage à deux phases est une technique aussi puissante que rentable. Neyman (1938) a été le premier à la proposer. Dans son ouvrage, et dans ses deux versions antérieures de 1953 et de 1963, Cochran (1977) présentait les résultats fondamentaux de l'échantillonnage à deux phases, y compris les estimateurs de régression les plus simples pour les plans d'échantillonnage de ce genre. Dans l'article que voici, nous adopterons un point de vue plus large et proposerons une approche générale à l'usage des données auxiliaires dans les plans d'échantillonnage à deux phases. Nous puiserons principalement pour cela dans les travaux de Särndal et Swensson (1987), de Särndal, Swensson et Wretman (1992) et de Dupont (1995). Des travaux plus récents en la matière comprennent ceux de Breidt et Fuller (1993), qui ont mis au point des méthodes d'estimation efficaces sur le plan des calculs pour l'échantillonnage à trois phases, en présence de données auxiliaires. Chaudhuri et Roy (1994) se sont pour leur part penchés sur les propriétés d'optimalité des estimateurs de régression plus simples mais bien connus de l'échantillonnage à deux phases. Enfin, Binder (1996) décrit une méthode simple de linéarisation permettant d'estimer la variance des estimateurs non linéaires. Cette méthode s'applique à tous les plans d'échantillonnage, notamment ceux à deux phases. Dans notre article, nous présumons l'emploi de plans d'échantillonnage *arbitraires* à chaque phase.

L'échantillonnage à une phase suppose l'emploi d'une couche d'informations pour l'estimation. Le nombre de couches double avec l'échantillonnage à deux phases, ce qui complique la tâche, car la manière idéale d'exploiter les données des deux sources n'est pas forcément évidente. Nous envisagerons deux approches à la construction d'estimateurs au moyen de données auxiliaires: l'*approche de régression généralisée* et l'*approche de calage*. Nous montrerons que la première ne constitue en réalité qu'une application particulière de la seconde. Les deux approches seront examinées dans le contexte de données auxiliaires à structure commune. On présume qu'on possède des données sur un vecteur auxiliaire x_1 couvrant les unités de la population entière et pour un deuxième vecteur auxiliaire x_2 couvrant les éléments de l'échantillon de la première phase. Au niveau de cette dernière, on dispose donc de données sur les deux vecteurs x_1 et x_2 . Särndal et coll. (1992) parlent de l'application de l'*approche de régression généralisée* à l'échantillonnage à deux phases. Ces auteurs élaborent un estimateur de régression généralisé pour l'échantillonnage à deux phases en attribuant un plan arbitraire à chaque phase de l'échantillonnage. Ils procèdent à deux ajustements par régression. Une régression «par le bas» engendre les valeurs prévues pour l'échantillon de la première phase, grâce aux données auxiliaires disponibles à ce niveau. Ensuite, une régression «par le haut» produit les valeurs prévues pour la totalité de la population au moyen de l'information appropriée. Les deux jeux de valeurs servent à bâtir l'estimateur de régression généralisé.

L'*approche par calage* met l'accent sur les poids attribués aux unités en vue de l'estimation. Le calage suppose la transformation d'un jeu de poids initiaux (habituellement ceux du plan d'échantillonnage) en

¹ M.A. Hidiroglou, Division des méthodes d'enquêtes-entreprises, Statistique Canada, Parc Tunney, Ottawa (Ontario) K1A 0T6; et C.-E. Särndal, Université de Montréal, et Statistique Canada.

- DEVILLE, J.-C., et SÄRNDAAL, C.-E. (1992). Calibration estimator in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- ISAKI, C.T., et FULLER, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.
- KOTT, P.S., et FETTER, M.J. (1997). A multi-phase sample design to co-ordinate surveys and limit response burden. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.
- LAVALLÉE, P., et HIDIROGLOU, M. (1988). Sur la stratification de population asymétriques. *Techniques d'enquête*, 14, 35-45.
- OHLSSON, E. (1995). Coordination of samples using permanent random numbers. Dans *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, et P.S. Kott). New York: Wiley, 153-169.
- SINGH, A.C., et MOHL, C.A. (1996). Comprendre les estimateurs de calage dans les enquêtes par échantillonnage. *Techniques d'enquête*, 22, 107-116.
- SIGMAN, R.S., et MONSOUR, N.J. (1995). Selecting samples from list frames of businesses. Dans *Business Survey Methods* (Eds. B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge, et P.S. Kott). New York: Wiley, 133-152.
- SKINNER, C.J. (1991). On the efficiency of raking ratio estimation for multiple frame surveys. *Journal of the American Statistical Association*, 86, 779-784.
- SKINNER, C.J., HOLMES, D.J., et HOLT, D. (1994). Multiple frame sampling for multivariate stratification. *Revue Internationale de Statistique*, 62, 3, 333-347.
- SWEET, E., et SIGMAN, R.S. (1995). *User Guide for the Generalized SAS Univariate Stratification Program*. Bureau of the Statistical Methods and Programming Division, Bureau of the Census, U.S. Department of Commerce, n° rapport ESM-9504.

que pour la variante de Poisson de la méthode fondée sur les NAP (où $\pi_{ij} = \pi_i \pi_j$).

Comme nous l'avons remarqué avec l'équation (4), l'erreur quadratique moyenne prévue de l'estimateur par calage est la même, que l'on utilise la méthode d'échantillonnage de Poisson fondée sur les NAP, la méthode NAP par collocation ou l'échantillonnage systématique PPT. Ces résultats portent à croire que l'estimateur de l'erreur quadratique moyenne par la méthode de Poisson pourrait être acceptable avec chacun des trois plans d'échantillonnage. Il existe, à l'appui de cette allégation, un autre argument plus solide axé sur le modèle, mais il ne sera pas invoqué ici.

7. DISCUSSION

Dans la section qui précède, nous avons indiqué que, si les poids de calage étaient conçus de manière à vérifier l'équation (2), alors l'estimateur obtenu serait sans biais dans le modèle de l'équation (3). Pour bon nombre d'applications, toutefois, il se pourrait qu'un autre modèle de calage convienne mieux que celui présenté dans l'équation (3). À titre d'exemple, si une variable de contrôle continue était utilisée pour la stratification d'une base de sondage particulière, il serait alors plus plausible d'utiliser cette variable directement dans le modèle, plutôt qu'indirectement par le biais des identificateurs de la base de sondage ou de la strate.

La méthode itérative du quotient est une forme de calage dans un modèle particulier. Il est donc tout indiqué d'utiliser le modèle le plus raisonnable qui existe. Par comparaison avec la méthode itérative du quotient, la méthode des moindres carrés à l'avantage de pouvoir être facilement appliquée aux variables de contrôle continues. Singh et Mohl (1996) présentent un examen détaillé d'autres algorithmes de calage, incluant une extension de la méthode itérative du quotient aux variables continues. Brewer (1994) propose une variante intéressante de la méthode des moindres carrés, qui ne figure pas dans Singh et Mohl (1996). Bon nombre des enquêtes économiques et agricoles utilisent des plans d'échantillonnage avec renouvellement, ce qui s'est avéré un moyen efficace d'assurer un équilibre entre les coûts et le fardeau de réponse. Bien que nos résultats empiriques privilégient l'échantillonnage systématique PPT pour les objectifs relatifs à la taille de l'échantillon, les trois plans d'échantillonnage fondés sur les NAP se prêtent beaucoup mieux au renouvellement de l'échantillon. (Voir par exemple Ohlsson (1995) à ce sujet.) Qui plus est, les méthodes fondées sur les NAP permettent d'intégrer différentes bases de sondage, à différentes périodes de l'année (avec l'échantillonnage systématique PPT, il est difficile de réaffecter l'échantillon à la base de sondage initiale). Ceci est une caractéristique particulièrement utile pour les enquêtes agricoles, du fait que les saisons de croissance varient d'une culture à l'autre.

BIBLIOGRAPHIE

- En résumé, le plan d'échantillonnage fondé sur les NAP avec échantillon fixe est excellent pour ce qui est d'atteindre les objectifs relatifs à la taille, mais il est difficile à utiliser en pratique parce que les probabilités de sélection sont habituellement inconnues et qu'elles doivent être simulées. Pour sa part, le plan d'échantillonnage systématique PPT est lui aussi très bon pour atteindre la taille visée, mais il est difficile à intégrer dans un plan avec renouvellement de l'échantillon. De plus, l'estimation de l'erreur quadratique moyenne requiert la formulation d'hypothèses relatives au modèle. Notre exemple empirique montre que l'échantillonnage par collocation donne des résultats qui ne sont que légèrement supérieurs à ceux obtenus par la méthode de Poisson, pour ce qui est des objectifs relatifs à la taille de l'échantillon. Il convient toutefois de préciser que des résultats différents pourraient être obtenus avec d'autres configurations des bases de sondage, des strates et des taux d'échantillonnage. L'échantillonnage par collocation se prête également aux plans avec renouvellement, comme l'échantillonnage de Poisson. Cependant, comme l'échantillonnage PPT, le premier requiert la formulation d'un modèle pour estimer l'erreur quadratique moyenne.
- Enfin, l'établissement de cibles pour p_{ij} ou n_{ij} constitue une méthode populaire, mais indirecte, de contrôler la variance de l'estimateur t_C associé à chaque base de sondage. Ce sont ces objectifs qui nous ont amené à prendre la décision ponctuelle de poser π_{ij} égal à $\max_j \{p_{ij}\}$. Une stratégie plus directe serait de définir des objectifs de variance prévus (asymptotiques) pour l'estimateur de chaque base de sondage, en utilisant l'équation (4) et les valeurs prévues pour $E_e(e_i^2)$. On pourrait alors choisir, disons la série de π_{ij} qui réduit au minimum la taille prévue de l'échantillon, tout en satisfaisant à ces objectifs de variance. Une approche similaire est adoptée par Amrhein, Fleming et Bailey (1997), qui utilisent l'algorithme de Chorny d'une manière analogue à Sigman et Monson (1995). La méthode d'échantillonnage de Poisson fondée sur les NAP, la méthode par collocation basée sur les NAP et l'échantillonnage systématique PPT demeurent trois solutions viables pour la sélection de l'échantillon, lorsque les π_{ij} optimales ont été déterminées.
- AMRHEIN, J.F., FLEMING, C.M., et BAILEY, J.T. (1997). Determining the probabilities of selection in a multivariate probability proportional to size sample design. Dans *Recueil Symposium 97: Nouvelles orientations pour les enquêtes et les recensements*, Statistique Canada, A paraître.
- BANKIER, M.D. (1986). Estimators based on several stratified samples with applications to multiple frame surveys. *Journal of the American Statistical Association*, 81, 1074-1079.
- BREWER, K.R.W. (1994). Survey sampling inference: some past perspectives and present prospects. *Pakistan Journal of Statistics*, 10(1)A, 213-233.

Supposons que $w'_0 = n_{(i)}/E[n_{(i)}]$ est le poids d'échantillonnage initial de l'unité i dans t_{ij}^N . De même, $w'_0 = 1/\max\{p_{ij}\}$ dans t_p^i et $1/\pi_i$ de façon plus générale, pour un estimateur de Horvitz-Thompson. Bankier (1986) propose d'utiliser la méthode itérative du quotient pour créer une série de poids corrigés, de sorte que

$$(2) \quad \sum_{i \in S_{jh}} w'_i = N_{jh}$$

pour chaque strate h dans chaque base de sondage f , où S_{jh} est la partie de l'échantillon qui se trouve dans la strate h de la base de sondage f , quelles que soient la ou les bases de sondage d où les unités ont été sélectionnées.

Deville et Särndal (1992) parlent d'une *équation de calage* (2). Ils soulignent qu'il existe diverses façons de calculer les poids de calage, w'_i , de manière à ce que l'équation (2) se vérifie et que la valeur de w'_i/w'_0 se rapproche, dans une certaine mesure, de 1 pour toutes les unités i . L'une de ces méthodes est la méthode itérative du quotient proposée par Bankier (1986). Une autre méthode, décrite en détail par Deville et Särndal (1992), est basée sur les moindres carrés. Cependant, que l'on utilise l'une ou l'autre méthode, nous obtenons l'estimateur

$$t_C = \sum_{i \in S} w'_i y_i,$$

où S représente l'échantillon complet, estimateur qui sera pratiquement sans biais puisque w'_i/w'_0 se rapproche de 1 pour toutes les unités i .

L'estimateur t_C est également sans biais sous le modèle:

$$(3) \quad y_i = \beta_0 + \sum_{f=1}^F d_{ifh} \beta_{fh} + \epsilon_i,$$

où la variable fictive, d_{ifh} , est égale à 1, lorsque l'unité i se trouve dans la strate h de la base de sondage f (échantillonnée ou non) et est égale à zéro dans les autres cas, alors que ϵ_i est une variable aléatoire avec une valeur moyenne de zéro. β_0 et β_{fh} sont des constantes inconnues (β_0 représente la valeur moyenne de y pour une unité dans la première strate de chaque base de sondage; c'est pourquoi la deuxième somme exclut $h = 1$). Les mêmes valeurs de d_{ifh} s'appliquent à chaque question d'enquête (y) d'intérêt, tandis que la valeur de β change pour chaque question. Pour bon nombre de questions, la valeur de β_{fh} sera égale à zéro lorsque la base de sondage f (disons les stocks de céréales) n'a aucun rapport avec la question (les superficiesensemencées avec de l'avoine, par exemple). Isaki et Fuller (1982) parlent de la valeur probable de l'erreur quadratique moyenne de t_C comme étant l'«erreur quadratique moyenne prévue» de l'estimateur. Cette valeur est très utilisée durant la phase de planification d'un sondage.

Si le modèle dans l'équation (3) se vérifie et que les ϵ_i sont sans corrélation, alors l'erreur quadratique moyenne prévue de t_C est

$$\begin{aligned} E[\text{EQM}_D(t_C)] &= E[E_D\{\sum_s w'_i y_i - \sum_p y_i^2\}] \\ &= E_D[E_D\{E[E_D\{\sum_s w'_i y_i - \sum_p y_i^2\}]] \\ &= E_D\{E_D[\sum_s (w'_i)^2 - 2w'_i E[E_D\{\sum_p y_i\}]] + \sum_p E[E_D\{\sum_p y_i^2\}]\} \\ &\approx E_D\{E_D[\sum_s [(1/\pi_i)^2 - 2/\pi_i] E[E_D\{\sum_p y_i\}]] + \sum_p E[E_D\{\sum_p y_i^2\}]\} \\ &= \sum_p [(1/\pi_i) - 1] E[E_D\{\sum_p y_i\}]. \end{aligned} \quad (4)$$

puisque $w'_i \approx 1/\pi_i$. Il est intéressant de souligner que, la valeur probable de la variance approximative du modèle, à la dernière ligne de l'équation (4), nous faisons la moyenne de tous les échantillons possibles et supprimons ainsi la principale source de variation entre les trois plans d'échantillonnage.

Supposons maintenant que nous avons utilisé l'échantillonnage aléatoire simple stratifié et que nous avons sélectionné l'unité i dont la probabilité est $p_{if} < \pi_i$, où f désigne la base de sondage qui se rapporte à y . On constate sans difficulté que la variance prévue de l'estimateur simple avec facteur d'extension aurait alors été $\sum_p (1/p_{if} - 1) E[E_D\{\sum_p y_i\}]$, ce qui donne une valeur pour le moins aussi élevée que le côté droit de l'équation (4). Il est donc avantageux – du moins pour les grands échantillons – d'intégrer les échantillons de diverses bases de sondage, comme nous l'avons fait ici. Nous ignorons toutefois quelle doit être, en pratique, la taille de l'échantillon pour que les résultats asymptotiques soient pertinents. Nous savons cependant que la taille de l'échantillon doit, tout au moins, être bien des fois supérieure au nombre de paramètres du modèle dans l'équation (3).

Il convient d'apporter quelques précisions sur l'estimation de l'erreur quadratique moyenne de t_C . L'estimateur de l'erreur quadratique moyenne, préconisé par Deville et Särndal (1992) – un estimateur qui allie à la fois un bon plan et des propriétés basées sur le modèle – ne peut être appliqué, à moins que la probabilité de sélection conjointe (π_{ij}) pour chaque paire d'unités d'échantillonnage (i et j) soit connue. Or, de tous les plans d'échantillonnage dont nous avons discuté, ces probabilités ne se calculent facilement

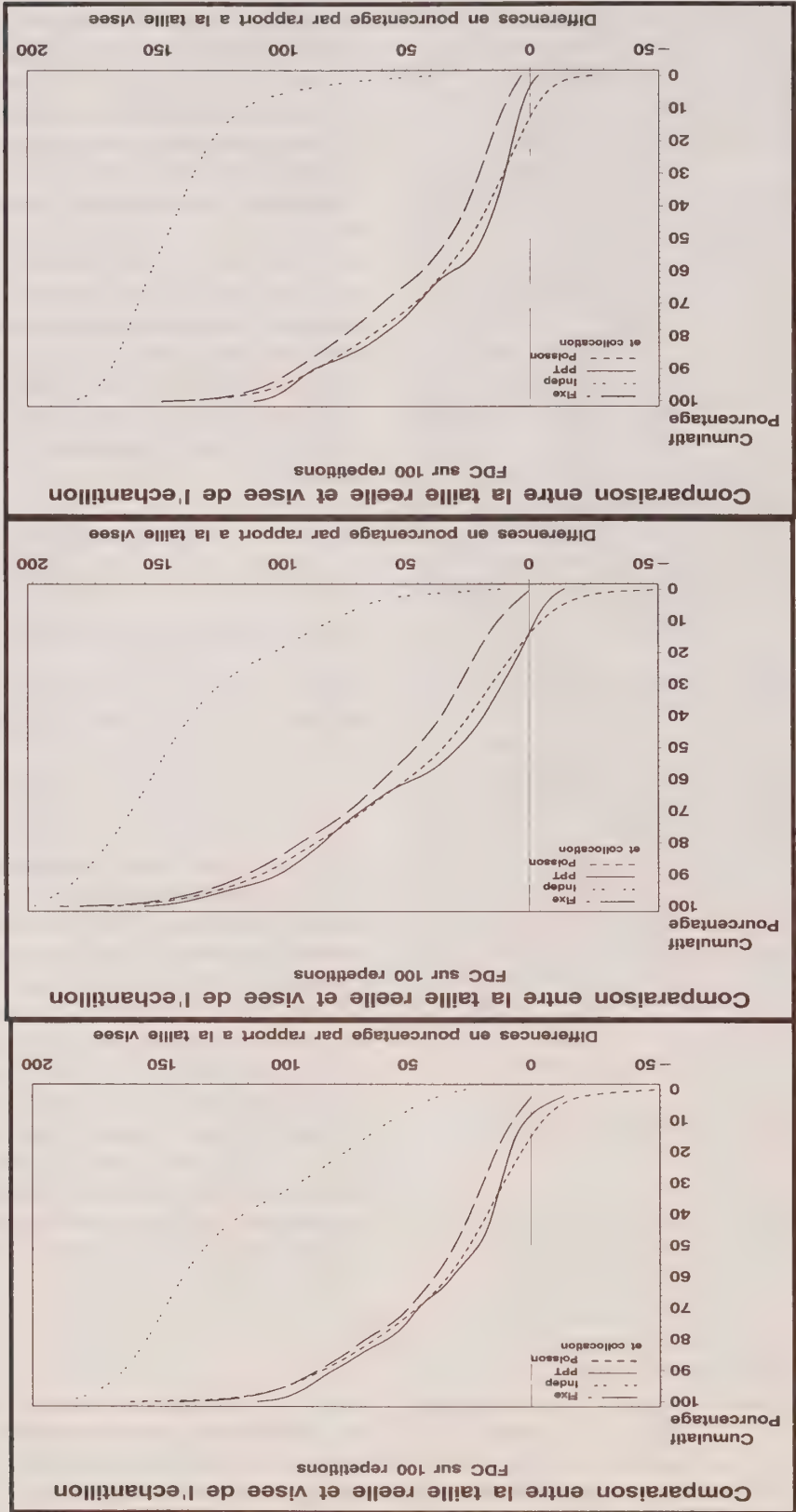


Figure 1. Comparaison entre la taille réelle et visée de l'échantillon pour les strates échantillonées. Ligne supérieure - MI; ligne médiane - CA; ligne inférieure - NJ

dans la base de sondage d'un produit particulier, mais qui se retrouve néanmoins dans l'échantillon du fait qu'elle a été sélectionnée dans une autre base de sondage. La présence de ces visiteurs a tendance à produire des échantillons dont la taille est en moyenne supérieure à la taille visée.

Tableau 1

Tailles réelles moyennes de l'échantillon (plus de 100 répétitions)

État	Base de sondage indépendante	Méthode avec échantillon fixe	Méthode NAP de Poisson	Méthode NAP par collocation	Méthode NAP par collocation systématique PPT
CA	496 (8,8)	388 (9,6)	375 (11,1)	374 (5,6)	373 (0,14)
MI	658 (9,3)	513 (9,2)	504 (13,6)	501 (6,0)	502 (0,48)
NJ	563 (8,1)	359 (8,6)	343 (13,8)	344 (4,6)	343 (0,17)

La taille de la population s'établit comme suit: CA-775; MI-1 041; NJ-785.

Les écarts-types sont indiqués entre parenthèses.

Tableau 2

Pourcentage des strates probabilistes pour lesquelles la taille réelle de l'échantillon est inférieure à l'objectif visé (100 répétitions)

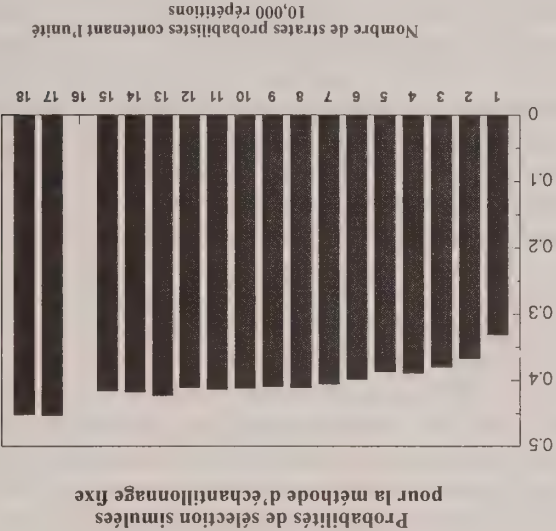
État	Méthode NAP de Poisson	Méthode NAP par collocation	Méthode NAP par collocation systématique PPT
CA	11 %	11 %	6,3 %
MI	12 %	12 %	6,3 %
NJ	11 %	8 %	1,4 %

La figure 1 (voir p.7) présente les distributions cumulatives des différences entre la taille réelle et la taille visée de l'échantillon; ces différences sont exprimées en pourcentages de la taille visée pour la strate échantillonnée; il s'agit, en d'autres mots, de la distribution cumulative de la taille (réelle-visée)/visée, au niveau de la strate probabiliste. Au Michigan, par exemple, il y a eu formation de 13 bases de sondage pour les denrées, chacune étant composée de deux strates probabilistes. L'échantillonnage à partir de ces bases de sondage a été répété 100 fois, de sorte que la fonction de distribution cumulative (FDC) pour chaque technique a utilisé 2 600 points. Les résultats obtenus avec les deux méthodes de Poisson sont représentés par une seule ligne, car ils coïncident. La méthode de Poisson ne produit pas un suréchantillonnage aussi marqué que ceux obtenus par les méthodes basées sur un échantillon fixe ou sur la base de sondage indépendante, mais elle comporte en revanche un risque de sous-échantillonnage, comme on le remarque au tableau 2. À l'inverse, les techniques basées sur un échantillon fixe (avec bases de sondage dépendante et indépendante) ne donnent pas lieu à un sous-échantillonnage, mais elles occasionnent davantage de suréchantillonnage que les méthodes de Poisson et PPT. Par ailleurs, la méthode PPT s'accompagne d'un certain sous-échantillonnage, qui n'est toutefois pas comparable à celui associé à la méthode de Poisson. Le plan d'échantillonnage PPT est également celui qui affiche la pente la plus prononcée de toutes les FDC, ce qui signifie qu'il y a moins de suréchantillonnage.

6. CALAGE

Le problème qui se pose, autant avec t_M qu'avec t_P (ou t_{HT} , est qu'ils ne sont souvent pas de très bons estimateurs de T en termes de précision (variance). Une des propriétés de l'échantillonnage aléatoire simple stratifié à base de sondage unique est que l'estimateur classique avec facteur d'extension estime parfaitement la taille de la strate (c.-à-d. avec une variance nulle). Or avec notre plan fondé sur des bases de sondage multiples, ni t_M , ni t_P ne pourra, dans la plupart des applications, estimer N_{jh} à la perfection.

Figure 2. Probabilités de sélection simulées pour la méthode d'échantillonnage fixe-Californie



Pareille simulation a été faite avec les données de la Californie. La technique avec échantillon fixe a été répétée 10 000 fois. Comme toutes les strates probabilistes ont été échantillonnées à un taux de 1/3, les probabilités simulées (c.-à-d. les fréquences relatives) peuvent être comparées à 1/3. Les probabilités de sélection simulées moyennes pour les 10 000 essais sont illustrées à la figure 2 par une fonction du nombre de bases de sondage dans lesquelles l'unité se trouve dans une strate probabiliste. Il existe 19 denrées d'intérêt dans cet État mais, dans exactement 16 ou 19 bases de sondage, il n'y avait aucune unité dans les strates probabilistes. Or la probabilité de sélection d'une unité a tendance à augmenter parallèlement au nombre de strates probabilistes qui la renferment. Cette probabilité de sélection est de 1/3 seulement lorsque l'unité est exacte-ment dans une de ces strates.

4. ÉCHANTILLONNAGE SYSTÉMATIQUE À LA TAILLE

Un autre plan d'échantillonnage, qui présente les mêmes probabilités de sélection que le plan d'échantillonnage de Poisson (et l'échantillonnage par collocation) décrit dans la section qui précède, se déroule comme suit:

- (0) Au besoin, créer une «strate» additionnelle pour chaque base de sondage, qui réunit les unités qui ne figurent dans aucune strate du plan d'échantillonnage.

- (1) Diviser la population en cellules s'excluant mutuellement, par classement croisé des strates de diverses bases de sondage. Une paire d'unités dans une cellule particulière se retrouvera ainsi dans la même strate de chaque base de sondage (p. ex. la grande strate des stocks d'avoine, la strate moyenne des stocks de céréales et la strate sans sorgho).

- (2) Classer par ordre aléatoire les unités dans chaque cellule, puis classer les cellules dans un ordre quelconque. On obtient ainsi une liste de toutes les unités de population.

- (3) À partir de cette liste, prélever un échantillon systématique avec probabilité proportionnelle à la taille (PPT), en utilisant la valeur π_i décrite dans la discussion sur l'échantillonnage de Poisson comme mesure de la «taille» (le mot «taille» apparaît ici entre guillemets, car π_i n'est pas, à proprement parler, une mesure de la taille). Cette approche assure que la probabilité de sélection d'une unité est égale à π_i .

Le plan d'échantillonnage systématique PPT décrit précédemment donnera toujours un échantillon dont la taille se rapproche de $\sum_{i \in P} \pi_i$. En fait, si $\sum_{i \in P} \pi_i$ est un nombre entier, alors la taille de l'échantillon sera exactement égale à cette somme. Dans les autres cas, la taille de l'échantillon sera égale à celui des deux nombres entiers qui se rapproche le plus de $\sum_{i \in P} \pi_i$. De même, le nombre prévu d'unités échantillonnées dans une cellule, C , sera égal à $\sum_{i \in C} \pi_i$, alors que la taille réelle de l'échantillon sera, soit $\sum_{i \in C} \pi_i$, soit un des deux nombres entiers qui s'en rapprochent le plus.

Examinons maintenant une strate particulière h dans la base de sondage f , pour laquelle la taille visée de l'échantillon est égale à n_{fh} . Pour une unité i dans cette strate, $\pi_i \geq n_{fh}/N_{fh}$ en vertu du plan. Supposons que $P(fh)$ désigne la série des unités de population dans la strate fh . Le nombre prévu d'unités échantillonnées dans fh est égal à $\sum_{i \in P(fh)} \pi_i \geq n_{fh}$. Rien ne garantit que la taille réelle de l'échantillon dans la strate sera supérieure ou égale à n_{fh} . Cependant, compte tenu de l'inégalité précitée et des limites inférieures de la taille de l'échantillon des cellules dans fh , la taille de l'échantillon dans la strate fh ne sera jamais de beaucoup inférieure à n_{fh} .

5. ÉVALUATION DES AUTRES TECHNIQUES D'ÉCHANTILLONNAGE

Les avantages de ce plan d'échantillonnage, par rapport à celui de Poisson ou à l'échantillonnage par collocation, tiennent au fait que ce plan produit un échantillon de taille plus stable et qu'il accroît la probabilité d'atteindre les exigences relatives à la base de sondage ou à la strate. Ces objectifs relatifs à la base de sondage ou à la strate sont en revanche toujours atteints avec l'échantillonnage fondé sur les NAP avec échantillon fixe, mais ceci à un prix: dans l'ensemble, la taille globale de l'échantillon est moins stable et les probabilités de sélection peuvent être très difficiles à déterminer.

Afin d'évaluer ces techniques d'échantillonnage empiriquement, nous avons choisi trois États dans lesquels est menée l'enquête sur l'utilisation de produits chimiques dans la production maraîchère du NASS, et nous avons répété 100 fois les trois techniques fondées sur les NAP, la méthode d'échantillonnage systématique PPT et l'échantillonnage indépendant, toutes bases de sondage confondues. Pour chaque répétition, les mêmes NAP ont été utilisés pour les trois techniques basées sur les NAP. Une base de sondage distincte a été créée pour chaque dentree à l'étude à l'intérieur d'un État (le nombre de bases de sondage varie de 2, au Minnesota, à 23 en Californie). Les unités de population ont été réparties comme suit entre deux strates probabilistes, une strate à tirage complet et une strate nulle. Les limites des strates ont été déterminées au moyen d'une méthode modifiée de Lavallée et Hidiroglou (1988), et les unités ont été réparties entre les strates selon la règle $\text{cum} \sqrt{f(x)}$ (Sweet et Sigman 1995). Cette stratification a été choisie, pour reproduire ce que serait un plan d'échantillonnage unidimensionnel acceptable ou assez répandu.

Un échantillon correspondant au tiers de la population a été prélevé de chacune des strates probabilistes. Le tableau 1 présente une comparaison de la taille globale de l'échantillon obtenu par chaque technique d'échantillonnage. Comme il fallait s'y attendre, le plan fondé sur la base de sondage indépendante a produit les échantillons les plus gros. Pour leur part, les trois techniques fondées sur les NAP ont donné des échantillons de taille similaire, la méthode de Poisson donnant les écarts-types les plus élevés lors de chacun des trois essais (États). Il semble que la méthode PPT soit la plus stable.

Le tableau 2 présente le pourcentage des échantillons prélevés des strates par la méthode de Poisson et la méthode PPT et dont la taille a été inférieure aux objectifs visés. Le fait qu'il n'y ait pas eu d'avantage de problèmes avec les tailles réelles par la méthode de Poisson s'explique notamment par la présence de ce que nous appelons les «visiteurs». Un visiteur est une unité qui n'a pas été choisie

réserverons pour la dernière section notre évaluation de la politique qui consiste à mettre l'accent sur des valeurs-cibles pour p_{ij}^h – ou encore pour n_{ij}^h . Nous nous contenterons ici de dire que bon nombre d'organismes statistiques, dont le NASS, ont une telle politique.

Selon un plan d'échantillonnage potentiel, un nombre aléatoire permanent (NAP) – obtenu de la distribution uniforme sur l'intervalle $[0, 1]$ – est attribué à chaque unité de la population. L'unité i est sélectionnée pour faire partie de l'échantillon de la base de sondage f , lorsque son NAP est inférieur à p_{ij}^h .

Nous obtenons ainsi un échantillon de Poisson, où la probabilité que l'unité i soit sélectionnée pour faire partie de l'échantillon correspond à $\pi_i = \max_f \{p_{ij}^h\}$, laquelle, manifestement, est au moins aussi grande que chaque p_{ij}^h individuelle pour une unité donnée. Selon un tel plan d'échantillonnage, l'estimateur sans biais de Horvitz-Thompson pour T correspond à $t_p^h = \sum_{i \in S} y_i / \max_f \{p_{ij}^h\}$.

Dans l'échantillonnage de Poisson, la taille de l'échantillon est aléatoire. Une façon de réduire la variance de la taille de l'échantillon est d'utiliser une variante de ce plan d'échantillonnage. Dans l'échantillonnage par collocation fondé sur les NAP, un NAP unique est attribué à chaque unité de population, lequel est choisi parmi les éléments de la série $\{e/N, (1 + e)/N, (2 + e)/N, \dots, (N - 1 + e)/N\}$, où e est une variable aléatoire uniforme prélevée dans l'intervalle $[0, 1]$. Pour ce faire, on peut d'abord tirer un NAP temporaire pour chaque unité, puis une valeur pour e . L'unité dont le NAP temporaire est le plus faible se voit attribuer un NAP par collocation dont la valeur est e/N , celle dont le NAP temporaire est le deuxième plus faible obtient la valeur $(1 + e)/N$, et ainsi de suite jusqu'à ce que la valeur $(N - 1 + e)/N$ soit assignée à l'unité dont le NAP temporaire est le plus élevé. L'estimateur t_p^h demeure sans biais dans l'échantillonnage par collocation.

En raison de la nature aléatoire de la taille des échantillons obtenus par échantillonnage de Poisson et relatifs à la taille de l'échantillon tiré de la base de sondage ou de la strate ne soient pas atteints pour un échantillon particulier que l'on tire. Un troisième plan d'échantillonnage fondé sur les NAP est mis en oeuvre avec comme objectifs les n_{ij}^h valeurs cibles, ce qui supprime cette possibilité. Dans ce dernier plan, les unités dans la strate h de la base de sondage f , affichant les n_{ij}^h plus faibles NAP, sont sélectionnées pour former l'échantillon (cette méthode est très similaire à l'échantillonnage séquentiel de Poisson utilisé par Ohlsson 1995). Avec ce plan fondé sur les NAP avec échantillon fixe, les probabilités de sélection des unités échantillonnées doivent être calculées pour l'estimateur de Horvitz-Thompson – une tâche difficile qui doit parfois être faite de façon approximative, par simulation.

pas nécessaire qu'au moins une base particulière soit complète. Il peut y aussi avoir chevauchement entre les bases de sondage.

Un estimateur sans biais pour la population totale $T = \sum_{i \in P} y_i$ est l'estimateur de multiplicité simple proposé par Skinner (1991):

$$(1) \quad t_M^h = \sum_{i \in P} y_i n_i^{(i)} / E[n_i^{(i)}],$$

où P représente l'ensemble de la population et $n_i^{(i)}$ est le nombre de fois que l'unité i est sélectionnée de quelque base de sondage que ce soit pour faire partie de l'échantillon. On remarquera que $n_i^{(i)} = 0$ pour les unités de population qui ne font pas partie de l'échantillon. Dans la grande majorité des applications, $n_i^{(i)}$ sera égal à 1 pour la plupart des unités échantillonnées, mais il est également possible que $n_i^{(i)} > 1$ avec ce plan d'échantillonnage.

Le nombre prévu de fois que l'unité i sera sélectionnée pour former l'échantillon correspond à $E[n_i^{(i)}] = \sum_f p_{ij}^h$, où p_{ij}^h est la probabilité de sélection de l'unité i dans l'échantillon aléatoire simple stratifié tiré de la base de sondage F , en d'autres mots, $p_{ij}^h = n_{ij}^h / N_{ij}^h$, où l'unité i se trouve dans la strate h , de la base de sondage f .

Il existe également un estimateur de Horvitz-Thompson pour T en vertu du plan, soit $t_{HT}^h = \sum_{i \in S} y_i / \pi_i$, où S représente l'échantillon et où $\pi_i = 1 - (1 - p_{i1}^h)(1 - p_{i2}^h) \dots (1 - p_{iJ}^h)$. Voir Bankier (1986) pour plus d'information sur cette approche.

3. STRATÉGIES D'ÉCHANTILLONNAGE UTILISANT DES NOMBRES ALÉATOIRES PERMANENTS

Le plan d'échantillonnage décrit précédemment est indépendant, toutes bases confondues. Pour bon nombre d'enquêtes, toutefois, il serait utile que le plan d'échantillonnage ne soit pas indépendant, car toutes les unités dans l'échantillon combiné ont le même instrument d'enquête et parce que bon nombre d'unités se retrouvent dans un certain nombre de bases de sondage. Par conséquent, étant donné les objectifs relatifs à la taille de l'échantillon prélevé de la base ou la strate, un plan d'échantillonnage qui aurait tendance à sélectionner la même unité dans chaque base de sondage devrait se traduire par un nombre moins élevé de contacts (et, partant, par des coûts d'enquête moindres) qu'un échantillonnage indépendant toutes bases confondues. Supposons, à cette fin, que chaque unité doit atteindre ou dépasser l'objectif p_{ij}^h dans chaque base de sondage. Cette valeur cible est constante pour toutes les unités qui se trouvent dans la strate h de la base de sondage f . Nous

Echantillonnage et estimation à partir de bases de sondage listes multiples

PHILLIP S. KOTT, JOHN F. AMRHEIN et SUSAN D. HICKS¹

RÉSUMÉ

Un grand nombre des enquêtes économiques et agricoles visent des objectifs multiples. Il serait donc pratique de pouvoir stratifier la population-cible de ces enquêtes de différentes manières – et ainsi répondre à un certain nombre d'objectifs – puis de combiner les échantillons pour le dénombrement. Nous examinons dans ce document quatre méthodes d'échantillonnage distinctes qui prélèvent des échantillons similaires, toutes stratifications confondues, ce qui permet de réduire la taille globale de l'échantillon. L'efficacité de ces stratégies d'échantillonnage est évaluée à la lumière des données extraites d'une enquête sur l'agriculture. Nous indiquons ensuite comment un estimateur par calage (c.-à-d. pondéré de nouveau) peut accroître l'efficacité statistique, en reproduisant dans l'estimation ce que l'on sait de la taille de la strate initiale. La méthode itérative du quotient, qui a été proposée dans certains ouvrages, n'est en fait qu'une méthode de calage.

MOTS CLÉS : Étalonnage; échantillonnage par collocation; nombre aléatoires permanents; échantillonnage de Poisson; échantillonnage systématique avec probabilité proportionnelle à la taille.

1. INTRODUCTION

Un grand nombre des enquêtes fondées sur une base de sondage liste, qui sont menées par le National Agricultural Statistics Service (NASS), sont intégrées, en ce que des données portant sur un éventail de sujets hétérogènes – depuis les superficies en culture aux stocks de céréales – sont recueillies par le biais d'une enquête unique plutôt que par l'exécution de plusieurs enquêtes indépendantes. Bankier (1986), Skinner (1991) et Skinner, Holmes et Holt (1994) ont démontré qu'il est possible d'accroître l'efficacité d'une vieille méthode qui consiste à combiner des échantillons aléatoires simples stratifiés prélevés séparément (où chaque échantillon provient d'une base de sondage liste dont le plan de stratification diffère); une telle stratégie d'estimation combinée donnerait des variances inférieures à celles qui seraient obtenues des enquêtes indépendantes compilées séparément.

Encore plus intéressant dans bien des cas serait un plan d'échantillonnage qui aurait tendance à sélectionner les mêmes unités de chaque base de sondage ce qui, par le fait même, réduirait les coûts et le fardeau de réponse associés à une enquête intégrée. Nous examinons ici plusieurs plans d'échantillonnage de ce type, dont trois sont fondés sur l'utilisation de nombres aléatoires permanents. Le quatrième plan utilise une variante de la méthode d'échantillonnage systématique avec probabilité proportionnelle à la taille. Le but visé, avec chacun des plans présents, est d'atteindre, voire de dépasser (du moins en moyenne) une série particulière d'objectifs relatifs à la taille des échantillons.

Nous démontrons également comment un estimateur par calage (pondéré) peut améliorer l'efficacité relative, en

reproduisant dans l'estimation ce que l'on sait sur la taille de la strate initiale. Dans la dernière section, nous montrons que l'utilisation d'une technique de calage peut servir à autre chose qu'à refléter uniquement la taille de la strate initiale.

Une autre stratégie pour réduire le fardeau de réponse consiste à utiliser des instruments séparés pour répondre à différents objectifs et à sélectionner des échantillons distincts pour chaque instrument. Cette méthode augmente le nombre total d'unités sélectionnées, mais elle réduit le fardeau de réponse pour chacune de ces unités. Le NASS utilise cette approche pour son étude sur la gestion des ressources agricoles (Agricultural Resources Management Study, voir Kott et Fetter 1997), mais ce n'est *pas* de cette approche dont il sera ici question.

2. ÉCHANTILLONNAGE INDÉPENDANT ET ESTIMATION SANS BIAIS

Supposons que nous ayons F bases de sondage indépendantes, par exemple une base de sondage pour les stocks de sorgho, une autre pour l'avoine et une base générale pour l'ensemble des stocks de céréales. Chaque base de sondage est stratifiée indépendamment et des échantillons aléatoires simples sans remise sont prélevés dans chaque strate, de chaque base de sondage. Supposons maintenant que la base de sondage f (disons celle de l'avoine) renferme H_f strates, que la strate h (grosses exploitations productrices d'avoine) dans la base f contient N_{fh} unités de population, dont n_{fh} unités sont sélectionnées. L'union de l'ensemble des F bases de sondage doit couvrir l'ensemble de la population (liste), mais il n'est

¹ Phillip S. Kott, Research Division; John F. Amrhein, Survey Sampling Branch et Susan D. Hicks, Estimates Division, National Agricultural Statistics Service, USDA.

Casady, Dorfman et Wang étudient la construction d'intervalles de confiance pour des paramètres de domaine lorsque la taille de l'échantillon du domaine n est pas déterminée par le plan de sondage. Ils rendent conditionnelle la taille de l'échantillon du domaine observé et montrent comment, dans certaines hypothèses relatives à la population, on peut obtenir des intervalles de confiance t conditionnels. Dans une étude empirique à partir de données tirées de l'enquête U.S. Bureau of Labor Statistics, Occupational Compensation, ils démontrent que les intervalles conditionnels proposés ont de meilleures probabilités de couverture que les intervalles marginaux ordinaires.

Montanari compare deux estimateurs bien connus de moyenne de population finie: ERG et l'estimateur de régression optimale du plan dérivé de l'estimateur par la différence. Alors que le premier est inefficace si le modèle sous-jacent est mal formulé, le deuxième, quoique indépendant d'un modèle, est sensible aux fluctuations d'échantillonnage. Par conséquent, une mesure d'efficacité comportant un critère de sélection d'un des deux estimateurs est prévue. Les résultats d'une étude empirique portant sur le comportement des deux estimateurs selon divers modèles mal formulés et exacts, sont analysés.

Haines et Pollock réexaminent les estimations de totaux à partir de bases de sondage multiples. Des estimateurs sont élaborés lorsque les données sont uniquement tirées de listes et lorsque les données proviennent également de bases aréolaires. Une simulation démontre que le meilleur estimateur est fonction de la dépendance connue ou présuée des bases. Haines et Pollock analysent également le cas où les observations sont disponibles pour toutes les unités ou pour un sous-échantillon de chaque base. À nouveau, le meilleur estimateur varie lorsqu'on tient compte de la dépendance entre les bases.

Bates et Gerber s'attaquent à la dynamique d'un problème difficile : le rôle joué par la mobilité temporaire d'une personne dans l'erreur de couverture à l'échelle du ménage. Ils élaborent une typologie à deux dimensions, puis, à partir de données tirées du Living Situation Survey menée aux États-Unis en 1993, ils dégagent quatre modèles de mobilité temporaire. Deux d'entre eux s'avèrent utiles comme prédicteurs de personnes omises dans les recensements ou les sondages.

Le rédacteur en chef

Dans ce numéro

Le numéro de *Techniques d'enquête* que voici renferme des articles sur des sujets variés. Kott, Amrhein et Hicks abordent la problématique des enquêtes à objectifs multiples. Pour de telles enquêtes, il serait souhaitable de pouvoir stratifier la population cible de diverses manières de façon à améliorer la précision des estimations d'intérêt. Les auteurs présentent quatre méthodes d'échantillonnage permettant de sélectionner des échantillons à travers les diverses stratifications tout en réduisant la taille globale de l'échantillon. Ces stratégies sont ensuite évaluées à l'aide de données provenant d'une enquête agricole. Ils montrent ensuite comment un estimateur par calage peut améliorer l'efficacité relative.

Singh, Horn et Yu examinent le problème de l'estimation de la variance de l'estimateur général de régression linéaire. Ils procèdent ainsi à un calage à deux niveaux distincts. Le calage à niveau élevé ainsi défini, recourt au total et à la variance connus des variables auxiliaires. Les auteurs montrent que cette méthode couvre une plus grande diversité d'estimateurs que l'approche de calage à bas niveau qui ne fait appel qu'au total connu des variables auxiliaires. Une étude empirique permet de juger de l'efficacité des stratégies proposées.

Hidiroglou et Särndal s'intéressent à l'emploi des données auxiliaires dans l'échantillonnage à deux phases. Ils présentent la façon dont ces données sont converties en poids de calage et ce, en deux étapes, dans le but de créer des estimateurs efficaces d'un total de population. Les auteurs montrent que l'estimateur de calage utilisant la fonction généralisée des moindres carrés peut être exprimé sous la forme d'un estimateur de régression à deux phases partiellement équivalent, c'est-à-dire d'un estimateur issu de deux ajustements par régression successifs. Ils examinent des formes de l'estimateur de calage à deux phases lorsque les données auxiliaires portent sur des sous-ensembles de la population appelés «groupes de calage». Ils discutent également de l'estimation de domaines d'intérêt et de l'estimation de la variance.

Byczkowski, Levy et Sweeney examinent les bases de sondage à structure multivoque, c'est-à-dire celles où toute unité de la base peut correspondre à des éléments multiples de la population cible et où tout élément de la population cible peut correspondre à des unités multiples de la base de sondage. Ce problème est soulevé par un sondage portant sur les caractéristiques des immeubles dans lequel la population cible est composée d'immeubles commerciaux mais dont la base est constituée d'une liste d'adresses (lesquelles correspondent à un seul immeuble, à plusieurs immeubles ou à des parties d'immeuble). Dans ce contexte, des estimateurs de totaux et de moyennes et leur variance ont été élaborés en employant un échantillonnage aléatoire simple et stratifié sans remise.

Yansaneh et Fuller présentent une méthode d'estimation de régression récursive permettant de diminuer la complexité du calcul liée à la meilleure estimation linéaire sans biais dans les cas d'enquêtes successives avec chevauchement partiel. À partir des données de la Current Population Survey (CPS) des États-Unis, ils comparent les variances de leur estimateur de régression récursive à d'autres estimateurs, y compris à l'estimateur composite du CPS. L'estimateur proposé semble être très efficace dans les estimations de niveau et de changement. Yansaneh et Fuller analysent également les variances suivant divers modèles de renouvellement et concluent que le modèle actuel de renouvellement 4-8-4 est supérieur au renouvellement continu dans le cas des moyennes actuelles de niveau et de longue période, mais inférieur dans le cas des changements sur de courtes périodes.

Lehtonen et Veijanen combinent deux idées bien connues, l'estimation de régression généralisée (ERG) et l'estimation du pseudo maximum de vraisemblance, pour élaborer une nouvelle méthode d'estimation du total d'une population pour une variable discrète d'enquête lorsqu'un vecteur de variables auxiliaires est connu. Les valeurs de la variable discrète sont modélisées en tant que réalisations d'une logistique multinomiale et les paramètres inconnus correspondants sont estimés selon un pseudo maximum de vraisemblance. Les fréquences de population à l'étude sont ensuite évaluées au moyen d'un estimateur ERG modifié qui comprend ces paramètres estimés. Les estimations de variance des fréquences sont obtenues au moyen d'une linéarisation par série de Taylor et certains résultats empiriques fondés sur l'Enquête sur la population active de la Finlande sont inclus.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 24, numéro 1, juin 1998

TABLE DES MATIÈRES

Dans ce numéro	1
P.S. KOTT, J.F. AMRHEIN et S.D. HICKS	
Échantillonnage et estimation à partir de bases de sondage listes multiples	3
M.A. HIDIROGLOU et C.-E. SÄRNDAL	
Emploi des données auxiliaires dans l'échantillonnage à deux phases	11
T.L. BYCZKOWSKI, M.S. LEVY et D.J. SWENEY	
Estimations à partir de bases de sondage de structure plusieurs à plusieurs	21
I.S. YANSANEH et W.A. FULLER	
Méthode optimale d'estimation réursive pour les enquêtes répétitives	33
S. SINGH, S. HORN et F. YU	
Estimation de la variance de l'estimateur général de régression: approche de calage à niveau élevé	43
R. LEHTONEN et A. VEIJANEN	
Estimateurs de régression généralisés logistiques	53
R.T. CASADY, A.H. DORFMAN et S. WANG	
Intervalle de confiance des paramètres de domaine quand la taille de l'échantillon du domaine est aléatoire	59
G.E. MONTANARI	
Estimation de la moyenne d'une population finie par régression	71
D.E. HAINES et K.H. POLLOCK	
Combinaison de bases multiples pour estimer la taille et les chiffres de la population	81
N. BATES et E.R. GERBER	
Mobilité temporaire et déclaration du lieu de résidence habituel	93

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président

G.J. Brackstone

Membres

D. Binder

G.J.C. Hole

F. Mayda (Directeur de la Production)

C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef

M.P. Singh, Statistique Canada

Rédacteurs associés

D.R. Bellhouse, University of Western Ontario

D. Binder, Statistique Canada

J.-C. Deville, INSEE

J.D. Drew, Statistique Canada

J. Eltinge, Texas A&M University

W.A. Fuller, Iowa State University

R.M. Groves, University of Maryland

M.A. Hidiroglou, Statistique Canada

D. Holt, Central Statistical Office, U.K.

G. Kalton, Westat, Inc.

R. Lachapelle, Statistique Canada

P. Lahiri, University of Nebraska-Lincoln

S. Linacre, Australian Bureau of Statistics

G. Nathan, Central Bureau of Statistics, Israel

Rédacteurs adjoints

J. Denis, P. Dick, H. Mantel et D. Stukel, Statistique Canada

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ par année au Canada et de 47 \$ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.



Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Juillet 1998

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

Publication autorisée par le ministre
responsable de Statistique Canada
© Ministre de l'Industrie, 1998

JUN 1998 • VOLUME 24 • NUMÉRO 1

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE





NUMÉRO 1

VOLUME 24

JUIN 1998

PAR STATISTIQUE CANADA

UNE REVUE
ÉDITÉE

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE

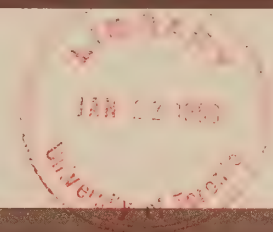


CA1
BS12
C001



Copyright
Public

SURVEY METHODOLOGY



Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1998

•

VOLUME 24

•

NUMBER 2



Statistics
Canada

Statistique
Canada

Canada



SURVEY METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

DECEMBER 1998 • VOLUME 24 • NUMBER 2

Published by authority of the Minister
responsible for Statistics Canada

© Minister of Industry, 1999

All rights reserved. No part of this publication may be reproduced,
stored in a retrieval system or transmitted in any form or by any
means, electronic, mechanical, photocopying, recording or otherwise
without prior written permission from Licence Services,
Marketing Division, Statistics Canada,
Ottawa, Ontario, Canada K1A 0T6.

January 1999

Catalogue no. 12-001-XPB

Frequency: Semi-annual

ISSN 0714-0045

Ottawa



Statistics
Canada

Statistique
Canada

Canada

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Survey Methodology is abstracted in The Survey Statistician, Statistical Theory and Methods Abstracts and SRM Database of Social Research Methodology, Erasmus University and is referenced in the Current Index to Statistics, and Journal Contents in Qualitative Methods.

MANAGEMENT BOARD

Chairman G.J. Brackstone

Members D. Binder
G.J.C. Hole
F. Mayda (Production Manager)
C. Patrick
R. Platek (Past Chairman)
D. Roy
M.P. Singh

EDITORIAL BOARD

Editor M.P. Singh, *Statistics Canada*

Associate Editors

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistics Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistics Canada*
J. Eltinge, *Texas A&M University*
W.A. Fuller, *Iowa State University*
R.M. Groves, *University of Maryland*
M.A. Hidioglou, *Statistics Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
R. Lachapelle, *Statistics Canada*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*

D. Pfeffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
F.J. Scheuren, *Ernst and Young, LLP*
J. Sedransk, *Case Western Reserve University*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R. Valliant, *Westat, Inc.*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

Assistant Editors P. Dick, H. Mantel, B. Quenneville and D. Stukel, *Statistics Canada*

EDITORIAL POLICY

Survey Methodology publishes articles dealing with various aspects of statistical development relevant to a statistical agency, such as design issues in the context of practical constraints, use of different data sources and collection techniques, total survey error, survey evaluation, research in survey methodology, time series analysis, seasonal adjustment, demographic studies, data integration, estimation and data analysis methods, and general survey systems development. The emphasis is placed on the development and evaluation of specific methodologies as applied to data collection or the data themselves. All papers will be refereed. However, the authors retain full responsibility for the contents of their papers and opinions expressed are not necessarily those of the Editorial Board or of Statistics Canada.

Submission of Manuscripts

Survey Methodology is published twice a year. Authors are invited to submit their manuscripts in either English or French to the Editor, Dr. M.P. Singh, Household Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6. Four nonreturnable copies of each manuscript prepared following the guidelines given in the Journal are requested.

Subscription Rates

The price of Survey Methodology (Catalogue no. 12-001-XPB) is \$47 per year in Canada and US \$47 per year Outside Canada. Subscription order should be sent to Statistics Canada, Operations and Integration Division, Circulation Management, 120 Parkdale Avenue, Ottawa, Ontario, Canada K1A 0T6 or by dialling (613) 951-7277 or 1 800 700-1033, by fax (613) 951-1584 or 1 800 889-9734 or by Internet: order@statcan.ca. A reduced price is available to members of the American Statistical Association, the International Association of Survey Statisticians, the American Association for Public Opinion Research, and the Statistical Society of Canada.

SURVEY METHODOLOGY

A Journal Published by Statistics Canada

Volume 24, Number 2, December 1998

CONTENTS

In This Issue	99
Longitudinal Surveys and Analysis	
Coordinating Editors Gad Nathan and Christopher Skinner	
D.A. BINDER	
Longitudinal Surveys: Why Are These Surveys Different From All Other Surveys?	101
F. BASSI, N. TORELLI and U. TRIVELLATO	
Data and Modelling Strategies in Estimating Labour Force Gross Flows Affected by Classification Errors	109
P.S. CLARKE and R.L. CHAMBERS	
Estimating Labour Force Gross Flows From Surveys Subject to Household-level Nonignorable Nonresponse ..	123
P.-A. SALAMIN	
Longitudinal Analysis of Swiss Labour Force Survey Data by Multivariate Logistic Regression	131
A.H. DORFMAN	
Price Index Surveys as Quasi-Longitudinal Studies	139
J.-L. TAMBAY, I. ŞCHIOPU-KRATINA, J. MAYDA, D. STUKEL and S. NADON	
Treatment of Nonresponse in Cycle Two of the National Population Health Survey	147
M.D. SINCLAIR and J.L. GASTWIRTH	
Estimates of the Errors in Classification in the Labour Force Survey and Their Effect on the Reported Unemployment Rate	157
R.H. RENSSSEN	
Use of Statistical Matching Techniques in Calibration Estimation	171
R. ARNAB	
Sampling on Two Occasions: Estimation of Population Total	185
E.L. KORN and B.I. GRAUBARD	
Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data	193
Acknowledgements	203

In This Issue

This issue of *Survey Methodology* begins with a special section entitled "Longitudinal Surveys and Analysis" which contains six of the papers presented at the IASS/IAOS Satellite Meeting on Longitudinal Studies held in Jerusalem in 1997. One or two other papers from that conference, which were not ready on time for this issue, may appear in future issues of the journal. I am very grateful to Gad Nathan and Christopher Skinner who were the Coordinating Editors for this special section. Without their persistence and hard work it would not have been possible.

The first paper in the special section, by Binder, introduces the topic by reviewing the current status and challenges for longitudinal studies as compared to cross-sectional studies. The discussion is divided into four parts, reviewing in turn the special issues and challenges encountered in the design, implementation, evaluation, and analysis of longitudinal surveys.

Bassi, Torelli and Trivellato consider the problem of estimation of gross flows among labour force states when there are classification errors in the data. They first review various strategies for the collection of longitudinal labour force data, and their likely implications for classification errors. They then present a general modeling framework and a modified LISREL model for adjusting gross flows estimates to correct for classification errors. The methods are illustrated by two case studies using data from the U.S. Survey of Income and Program Participation and the French Labour Force Survey.

Clarke and Chambers consider the impact of household level non-response on estimates of labour force gross flows. They propose a class of models for nonignorable household-level nonresponse. They then use simulations to demonstrate that labour force gross flows estimates can be biased in the presence of this nonignorable household level nonresponse, and that estimates using household level nonresponse models can reduce this bias. If the household level nonresponse mechanism is correctly specified then this source of bias is removed completely; however, even incorrectly specified household nonresponse models can reduce the bias.

Salamin considers the problem of estimating a change in proportion for a small area. He shows how a general multivariate logistic regression model can be used to describe the longitudinal data obtained from a rotating panel design. He also considers how the parameters of this model may be restricted to describe various types of dependence among the repeated observations, leading to alternative model based estimates of change. The method is illustrated by estimating changes in probability of being employed for a Canton in Switzerland using data from the Swiss Labour Force Survey. Compared to simple differences of estimated proportions of employed persons, the model based estimates have smaller standard errors.

Dorfman, in his paper, attempts to treat consumer price indices from a statistical point of view. He first reviews price index theory in general, including the stochastic approach and objections to it. He then proposes a modification to the stochastic approach, based on state space modeling, which circumvents the major criticism of it. The approach is illustrated using price and quantity data for canned tuna.

In the last paper in the special section, Tambay, Schiopu-Kratina, Mayda, Stukel and Nadon describe the treatment of nonresponse in the Canadian National Population Health Survey. Data collected at the first cycle of the survey are considered as potential predictors of nonresponse to the second cycle. A CHAID (Chi-square Automatic Interaction Detection) algorithm is used to determine weighting classes for nonresponse adjustment at the second cycle. The paper also briefly describes the sample design and other steps in the derivation of the estimation weights.

Sinclair and Gastwirth study the problem of misclassification error of labour force status in the Current Population Survey of the U.S. Bureau of the Census. To do so, they extend the method of Hui and Walter, which is appropriate for dichotomous data using reinterview data, to the trichotomous case. Unlike other methods, this method does not assume that reinterview data is error free, but rather assumes an error in both the original interview and the reinterview data. They make an empirical assessment by comparing the estimated error rates generated by their method as opposed to other existing methods such as that of Poterba and Summers, and find that the degree of underestimation of the error tends to be higher when the true unemployment rate is in fact high. Finally, rather than assuming a constant error rate throughout, they attempt an analysis assuming that the error rates are constant only within time groupings having differing levels of unemployment.

Renssen considers the problem of combining information on variables collected from two different large surveys, using auxiliary information from a smaller third survey collecting all of the variables. Using ideas from statistical matching and from calibration, he proposes methods for the production of two-way tables, for the production of microdata files, and for the estimation of correlations. For the production of two-way tables his development leads to consideration of two different sets of calibration constraints, one termed incomplete two-way stratification and the second termed synthetic two-way stratification. In a simulation study using data from a pilot study for the Dutch Household Survey on Living Conditions, the calibration based on synthetic two-way stratification is shown to be much better.

Arnab considers different strategies for sampling on two occasions. The sample at the second occasion is assumed to be a combination of a subsample of the first sample and a new, unmatched sample. Different strategies for subsampling the first sample and estimating a total at the second occasion are compared. He reviews strategies already existing in the literature, and proposes two new ones. Efficiencies of various strategies are compared analytically and empirically.

Finally, Korn and Graubard consider the problem of generating confidence intervals for proportions having a small expected number of positive counts. Noting that the Clopper-Pearson binomial intervals traditionally used in the non-survey setting are inappropriate for use with complex survey data, they propose a modification of these intervals. Via simulation, they then compare the proposed intervals to others commonly used such as: logit-transform intervals, Breeze (1990) intervals based on a Poisson approximation, and normality-based linear intervals. They also illustrate the proposed and three alternative methods with applications using data from both the National Health and Nutrition Examination Survey and the Hispanic Health and Nutrition Examination Survey.

The Editor

Longitudinal Surveys: Why Are These Surveys Different From All Other Surveys?

DAVID A. BINDER¹

ABSTRACT

We review the current status of various aspects of the design and analysis of studies where the same units are investigated at several points in time. These studies include longitudinal surveys, and longitudinal analyses of retrospective studies and of administrative or census data. The major focus is the special problems posed by the longitudinal nature of the study. We discuss four of the major components of longitudinal studies in general; namely, Design, Implementation, Evaluation and Analysis. Each of these components requires special considerations when planning a longitudinal study. Some issues relating to the longitudinal nature of the studies are: concepts and definitions, frames, sampling, data collection, nonresponse treatment, imputation, estimation, data validation, data analysis and dissemination. Assuming familiarity with the basic requirements for conducting a cross-sectional survey, we highlight the issues and problems that become apparent for many longitudinal studies.

KEY WORDS: Frames; Administrative data; Data collection; Nonresponse; Imputation; Estimation; Data analysis.

1. REASONS FOR LONGITUDINAL STUDIES

Each year around the world various statistical agencies conduct thousands of surveys. Usually, these surveys obtain information required for decision or policy making. These surveys are not conducted just for historical purposes, but also to have information on what measures may be taken to assist with making various policy changes. Most surveys are based on cross-sectional data, where a survey is taken of a particular population at a given point in time. Various summaries are taken about the population under consideration at the time of the survey. However, very often the interest is not so much in what actually happened when the survey was taken, but what would be the impact of making various changes. Alternatively, a planned change in policy may be forthcoming and monitoring the effect of this change is desirable. What is most important is the time element. For example, when trying to learn about certain phenomena such as health status or education attainment, one is interested in the various determinants related to these outcomes. Sometimes, the actual temporal relationship is not even clear in terms of what are the causes that precede the effects. These could be measured if, instead of taking a cross-sectional survey, surveys are conducted over time, either as a series of cross-sectional surveys or, alternatively, using the same panel of respondents from one occasion to another. This common sense notion has led to the desire to conduct more longitudinal studies. This also has the benefit that the effects of unobserved variables may be less important when the same respondents are used to compare differences over time.

One of the factors contributing to the increase in the number of longitudinal studies is that administrative data

sources can now be used more effectively, thus making certain longitudinal studies feasible. Administrative data are becoming increasingly available. These data are often routinely collected for the same individuals over a period of time. Even if the data collected from the administrative sources is not ideal for the survey-taker, they may provide a good proxy for the information.

The advantage of designing a study as longitudinal is that a common methodology can be used for each of the various waves of the survey. This may lead to more valid conclusions. Often, when trying to understand various patterns of social and economic change, conducting surveys of the same respondents on a number of occasions is best. Less desirable, but possibly satisfactory, is simply to repeat the survey from one occasion to another without necessarily returning to the same respondents. This may be less costly. The main point is that to understand certain phenomena over time, collecting the information on more than one occasion is necessary.

When making decisions on the nature of a new longitudinal study, a number of cost considerations need to be accounted for. Obviously, one needs to consider the benefits against these various costs. Issues that longitudinal studies could address cover many subject-matter areas. We enumerate just a few of them. In the area of health status, one is interested in changes to health status and the determinants that lead to these changes. In other words, what are the health risks, and what, in fact, is the effect of these health risks on health status in the long term? By collecting the data from the same individuals over a period of time, one can assess these factors, not just on small scale studies typical of clinical trials, but on large-scale nationally-based population health surveys. However, the type of information that can be obtained from a nationally-

¹ David A. Binder, Business Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

based longitudinal survey would be very different from that which is obtainable in a clinical trial.

Another topic where there is interest in observations over time is in the area of labour and income. For example, it is not enough to have information on the net change to labour force status and labour force participation rate over time. It is also of interest to know which individuals move from say, being unemployed to working or to not being in the labour force. In recent times, employment patterns have changed. More women are working and part-time work is more common. Frequency of job changes is also changing. To understand these phenomena, longitudinal surveys can answer many important questions. The characteristics, for example, of entry level jobs taken by those who were previously unemployed may be of interest, as well as effectiveness of different job search strategies by individuals or the effectiveness of various government training schemes.

Length of spells in poverty is of increasing interest. For example, for persons with low income, how long does one remain in that situation? What are the various factors that will determine whether this is a long-term situation? How important are education and other factors with respect to poverty and the length of poverty spells?

In the field of education, an interesting aspect is the school-to-work transition at the time when people finish full-time school and decide to join the labour force. This behaviour may be measured more easily through a longitudinal study than through other types of surveys. Another education-related example is the effectiveness of various types of education such as vocational training and adult training programs.

In justice and victimization, there are many examples where observing the same individuals over time can be beneficial. Persons who have been victimized could be followed up to assess the long-term implications. As well, persons who have been involved with the judicial system may be observed over time to determine the subsequent patterns of behaviour and the determinants for these patterns.

Studies of consumer behaviour are of great interest to marketers and others. This would include purchasing patterns for consumers. Event histories for consumer purchasing would be very useful to many researchers.

Studies on the effects of government transfer payments to individuals over time can be important to policy makers. A longitudinal study can determine how long individuals may be dependent on such government payments, whether or not habits are created because of the existence of some of these payments, what are the characteristics of the individuals and what are the long-term effects of participation in various assistance programs.

On the economic side, the longitudinal characteristics of various businesses are of great interest. One can measure how efficient these businesses are, what the use of technology is in these businesses, what is the long-term

effect of this use and how productivity is changing over time. Various interesting questions on business demographics could be asked; for example, what are the characteristics of businesses that result in failure, what are the economic conditions under which businesses are created. As well, mergers and amalgamations are of interest with respect to the conditions under which these occur. Through longitudinal studies these phenomena can be more easily measured.

There have been various structural changes to many businesses over the last few years and it is only through longitudinal studies that one can observe some of these structural changes at the micro level. Many measures can be estimated only when the respondents are measured on more than one occasion.

Another area of interest is in agriculture, where the nature of farming is undergoing transition. Of interest is how farms are changing, both in terms of the products that are being produced and the size of the farms. Changes in the characteristics of who is running the operation are also of interest.

As we have discussed, there are many applications and many facets to longitudinal studies. Also, there are many dimensions to their design and analysis. In the following sections we summarize these issues around Four Questions: design issues, implementation issues, evaluation issues and analysis issues. Many of these issues have been discussed in Kasprzyk, Duncan, Kalton and Singh (1989) and in Armstrong, Darcovich and Lavallée (1993). Some design issues and time series methods are reviewed in Binder and Hidioglou (1988). We include a few more recent references.

QUESTION 1: DESIGN ISSUES

When designing a longitudinal study, advance planning is vital to the success of the study. For example, one must ensure that only relevant and accurate information is being collected from the respondents so that the potential benefit of the longitudinal survey is maximized. This implies that the longitudinal analyses to be undertaken from the survey should be planned from the outset to ensure that the relevant data are obtained. Duncan and Kalton (1987) give an excellent summary of many of the issues. Webber (1994) describes the testing strategy used in the planning of the Survey on Labour and Income Dynamics. Huggins and Fischer (1994) discuss the plans for the redesign of the Survey of Income and Program Participation based on their experiences. Longitudinal studies can be more expensive than a series of cross-sectional studies. Therefore, the benefits of collecting these data must be even greater since the costs themselves are higher. As well, ensuring that funding for a longitudinal study can be assured is important since the fruits from the longitudinal nature of the study may not be borne until at least the second or third wave of

the study. There is a difference, of course, between planning for a study to be longitudinal from the beginning as opposed to taking a series of cross-sectional data and trying to merge them into a longitudinal database. Obviously, the former is more desirable but often, because of the history of the survey-taking organization, a series of cross-sectional data already exists so that merging these would be a reasonable alternative; see Hughes and Hinkins (1995).

In general, careful attention needs to be paid to the design of the database for any longitudinal survey where the analysis includes longitudinal measures such as the study of episodes and spells. For some statistical agencies and organizations, the survey program is now in transition from cross-sectional surveys to longitudinal surveys. The change from a series of cross-sectional surveys to longitudinal surveys requires careful planning. When conducting longitudinal surveys, the databases need to be maintained and updated in ways that are very different from cross-sectional surveys. There may be many infrastructure and organizational issues within the agency that become apparent as more longitudinal surveys are being conducted, particularly with respect to the maintenance of the databases and the survey operations. The impact of such changes on the statistical organization may be substantial.

An important issue to consider when planning for a longitudinal survey is whether or not the users will also be requiring cross-sectional estimates. Is there a requirement to have information about the respondents who are in the survey over a period of time, and also being able to produce estimates for a single point in time as if it were a cross-sectional survey? If this is the case, there are major implications on the way the survey is designed and implemented; see Lavallée (1995). This concern would also be present if the variables of interest include comparing cross-sectional estimates over time, as opposed to true longitudinal measures such as studying autocorrelations for common units in a business survey.

Concepts and the definitions used in longitudinal surveys are usually obtained through consultations with the data users. Even the definition of the longitudinal unit to be observed over time may need clarification for dynamic populations. This is the case for both household surveys and for business surveys. Understanding the user requirements and discussing what can be measured over time with appropriate quality is important. During the survey planning, these requirements must be carefully weighed against what is operationally feasible in an actual survey context. Given the eventual costs of these studies, conducting thorough tests is often worthwhile, particularly on the survey questionnaires. A point that deserves more attention is the need for more standard longitudinal measures that are common across countries. This would permit governments and researchers to make better international comparisons.

Another major component for designing longitudinal studies is the creation, use and maintenance of sampling

frames over time in ways that facilitate the implementation of the study. For example, an establishment panel survey may be based on a business register that can be highly dynamic with respect to births, deaths, mergers and amalgamations. It is important that the definitions of which units are to be included in these panels over time are clear under these conditions.

One reason that longitudinal surveys have become more prevalent in recent years is the fact that there are more administrative data files available now that can be used as frames for conducting the longitudinal studies. The administrative files themselves may also contain useful data information besides just being useful as frames per se. Some data manipulation of the administrative data is usually required to make these data useful for the statistical purpose of the longitudinal study, however. In general, the impact of frame changes to the study must be carefully considered at the design stages.

A common practice is to take a number of different administrative files and to match them to create a sampling frame. As well, some longitudinal studies are based solely on the information contained in various administrative files. The difficulty, of course, is that over time these administrative files will change. This may imply a change to the samples that are being taken from these files, and therefore special measures will need to be taken to keep the analyses relevant.

Often a longitudinal study is based on an existing survey or census conducted at a point in time in the past, and this then becomes the basis for the sampling frame for following up respondents over time. One disadvantage of this is that it becomes difficult to obtain cross-sectional estimates when births to the population are excluded from the frame. Record linkage techniques may be necessary for maintaining the frame and such techniques are usually error-prone.

For rare populations, it is often advantageous to use not just a single frame but to use multiple frame methods. This ensures that there is adequate representation from the populations of interest that might be underrepresented in a single frame, but this may also require the use of record linkage and complex weighting techniques.

An important design issue is the method of sampling from the frame once it has been established. In Kalton and Citro (1993), a number of different types of longitudinal surveys were enumerated. These were repeated surveys, that is, a series of cross-sectional surveys; panel surveys, where certain respondents are selected and followed up over time; repeated panel surveys, where new panel surveys are selected at different points in time; rotating panel surveys, where on each occasion a panel is dropped from the study and a new panel is added; overlapping surveys, where there are common respondents from one occasion to the other, but not necessarily through a fixed panel sample design; split panel surveys that can be a combination of panel surveys and repeated or rotating panel surveys. The

sample design must ensure that there is a sufficient sample from the population of interest as well from any of the control groups. Administrative data have proven to be very useful when designing a sample for many of these surveys as they often provide a suitable frame.

As a referee pointed out, a key issue at the design stage is the strategy for dealing with sample loss through attrition, due to nonresponse, leaving the target population, *etc.* Possibilities include topping up the sample in subsequent waves, but such a strategy can distort the representativity of the cohort. Another strategy would be to start with a larger sample and not replace lost units; see, for example Singh, Petroni and Allen (1994).

When deciding on a particular sample design, consideration must be given to the related weighting and estimation issues. As well, the periodicity or frequency of the survey must be established. Obviously, when the variables of interest change more rapidly, having the survey conducted more frequently would be more desirable. On the one hand, more frequent surveys lead to increased cost and respondent burden; on the other hand, less frequent surveys can lead to larger recall biases. These cost-quality tradeoffs are usually difficult to quantify.

Very often, if both cross-sectional and longitudinal estimates are required, ensuring that there will be valid cross-sectional estimates may be necessary to select supplementary samples. This is because there may be members of the population in the cross-sectional estimates who were not in the sampling frame on previous waves and, therefore, would not be represented in the sample. Czajka (1994) studies this for the case of estimating income.

Designing some evaluation samples is also worthwhile at the planning stage. There are a number of sources of bias in longitudinal surveys. Some of these biases can occur simply because the same respondent has been surveyed on a number of occasions. Therefore, consideration should be given to adding additional samples for evaluation purposes only, in order to be able to measure some of these impacts. These samples would include individuals in the target population that were not in the longitudinal survey. They are most useful for evaluating cross-sectional measures.

QUESTION 2: IMPLEMENTATION ISSUES

The second main issue we discuss is related to the implementation of a longitudinal study. First, one has various choices of modes of data collection. Recently, computer-assisted interviewing has gained popularity. With computer-assisted interviewing, more choices of survey instruments are available. For example, using dependent interviewing where the respondent or the interviewer has access to the responses from previous occasions is easier. This may increase or decrease certain biases. Hill (1994) assesses this in the context of Survey of Income and Program Participation.

Of course, since we are going back to the same respondents on a number of occasions, the question of response burden is even more crucial than in a single cross-sectional survey. We do not want to overload the respondent since this could result in higher refusal rates at later waves of the survey. Michaud, Dolson, Adams and Renaud (1995) suggest respondent burden can be reduced by making more use of administrative data. Reducing attrition due to nonresponse is an important goal in longitudinal surveys and consideration may be given to the use of monetary or other incentives to help keep the integrity of the sample over time; see Lengacher, Sullivan, Couper and Groves (1995). Another means of reducing attrition is to collect information to aid in the tracing efforts and to keep in contact with the respondents over time; McGuigan, Ellickson, Hays and Bell (1995) studies alternatives of tracing, reweighting and sample selection modelling, to cope with attrition problems.

In some longitudinal surveys, some data are collected retrospectively; that is, questions are asked which refer to previous points in time as well as the current point in time. This could lead to what is known as seam effects. As a result, the observed changes over the reference periods may depend on which periods contain data obtained retrospectively.

Administrative records may be useful to enrich the database so that not all data need to be collected directly from the respondent; see Michaud *et al.* (1995). Of course, this could depend on the quality of the administrative data, its availability, and what the interplay is between the information from the administrative records and the survey variables; see Stearns, Kovar, Hayes and Koch (1996) for an example that studies this relationship. When dealing with administrative data or merged sample files, there may be data gaps in these various files and how to handle these data gaps becomes an issue.

In general, changes to the frame structure can result in difficulties when performing the longitudinal analyses. Some key characteristics of the respondents could also be changing over time. For example, in a business register, if the industrial classification information changes because of the fact that businesses change the nature of the products that they are producing over time, being able to keep track of this changing classification on the database to ensure that the longitudinal analyses are as useful as possible is important. This can also complicate the analysis.

Many issues arise when the database is obtained by combining the samples from a series of individual surveys. Integrating this information may present a challenge because different surveys may have used different methodologies. This could result in some inconsistencies in the quality of the information from one database to another.

Important issues for many longitudinal surveys are those related to record linkage. Record linkage is used in many processing steps. In some cases, the longitudinal studies may be based solely on these linked files. Record linkage

is common for creating and maintaining the survey frames, including linking administrative files over time, linking administrative files and survey frames and linking separate survey frames. For example, for surveys of establishments, we may wish to create longitudinal composite records for the establishments that are based on several independent repeated surveys, since many of the establishments are surveyed on each occasion. Record linkage is often used to find which units correspond to the same establishments. Record linkage is also used to identify births to a frame. Of course, the errors due to the record linkage can be important in the analysis; see Scheuren and Winkler (1993).

In some cases, in fact, no real respondents are being followed over time. Instead record linkage is used to create artificial populations through statistical matching. These populations are then analysed as if they were real.

Another implementation issue is that of handling non-response. It is known that nonresponse to longitudinal surveys does not occur completely at random. There tends to be differential nonresponse among different subpopulations. Therefore, special attention needs to be placed on how the imputations or reweighting will be performed; see, for example, Tambay, Şchiopu-Kratina, Mayda, Stukel and Nadon (1998). When using administrative data as the basis for the longitudinal study, there may be missing administrative data and special procedures will be necessary to handle this situation.

For missing data, there are generally two methods of treatment: imputation and reweighting. Reweighting is common for situations where there is wave nonresponse. Imputation is more frequently used when there is partial nonresponse within a given wave of the survey. There can be advantages to longitudinal imputation as opposed to cross-sectional. For longitudinal imputation, the longitudinal information from the same individual on the database is used as the basis for doing the imputation, as opposed to using other individuals at the same point in time. For attrition and wave nonresponse, one may wish to model the attrition rates and use these models to compensate for the nonresponse through weight adjustments. A variety of weight adjustments were researched for the Survey of Income and Program Participation and the results were presented in Rizzo, Kalton and Brick (1994), Folsom and Witt (1994), and An, Breidt and Fuller (1994). Singh, Wu and Boyer (1995) study this problem for the difficult case of estimating gross flows.

There are many complexities that may be introduced into the derivation of the weights. There are various approaches and techniques available to calculate both cross-sectional weights and longitudinal weights. Cross-sectional weights are used for measures of the population at a single point in time, whereas the longitudinal weights are necessary when data from individuals over more than one occasion are included. The analyst may wish to have person-level weights that are different from the household-level weights; Kalton and Brick (1995). For example, for some variables

such as household income, using household-level weights would be preferable to the individual person-level weights. Weighting becomes more complex with the use of multiple frames. Effective use of administrative data may imply even more complexities in the weighting scheme itself; see, for example Stearns *et al.* (1996).

There are many causes for the samples to become unrepresentative. For example, lack of representativity could be due to problems of coverage due to immigration into the population. Some undercoverage may be due to attrition. Some overcoverage could be due to including some non-sampled co-habitants of a household, thus implying that those individuals could be included in the sample by living with an originally sampled person; see Lavallée (1995) and Kalton and Brick (1995). Other types of systemic overcoverage are also possible. Ensuring that no biases are introduced requires special weighting treatments. For longitudinal surveys in particular, this may become quite complex. Administrative data can be used both to assess whether or not the sample is representative and to provide information for making the appropriate adjustments.

Since much of the estimation for longitudinal study will be associated with measuring change as opposed to measuring the phenomena at a single point in time, there will be questions about how to develop the variances for these estimates of measures of change. Some new procedures may need to be developed for this situation. In general, variance estimates can become quite complicated when the statistics are complex functions of the longitudinal observations. For example, income class boundaries may change over time and studying the transitions of individuals from one class to another is of interest.

Another complexity of estimation may be the desire to include information from ongoing cross-sectional surveys to produce new integrated measures, using all the information that is available from the various available sources.

QUESTION 3: EVALUATION ISSUES

The third set of issues we discuss is related to the evaluation of the information and methods. Even though the evaluations may be conducted separately from the implementation, the results of such evaluations should impact on the survey itself, either by altering the estimation methods or by changing the way the survey is designed and implemented in future waves.

There are many sources of biases that could be studied. Biases may be due to dependent interviewing by giving the respondent and the interviewer information that could refer to a previous occasion of the survey. Seam effects can arise from retrospective studies; see, for example Murray, Michaud, Egan and Lemaître (1991). Other sources of bias could occur when the nonresponse is informative; that is,

when the nonresponse propensity is related to the variable of interest. An example would be when household level nonresponse is correlated with gross flows within the household, where gross flows are the changes in the individual's classification; see Clarke and Chambers (1989). Other biases could be due to measurement or classification errors; see, for example, Bassi, Torelli and Trivellato (1998). Conditioning bias could arise from the fact that since we have been asking the respondents about information, such as labour dynamics, they may have become more sensitized to some of these issues so that their behaviour could change because of the fact that they are included in the survey.

The effect of response errors and interviewer errors on the analysis should be evaluated. Different individual interviewer methods may lead to different error rates. The stability or instability of the turnover of interviewing staff could affect some analyses. Questions such as whether or not the information was collected by proxy can also be relevant.

Other evaluations could be performed to measure the effect of attrition and to evaluate various imputation methodologies and other nonresponse handling strategies; see Tin (1996) for an evaluation of attrition using econometric methods. Schejbal and Lavrakas (1995) study the effect of panel attrition in a dual-frame local telephone survey. Corder, Manton and Woodbury (1994) study ways to improve coverage and reduce attrition in the context of the National Long Term Care Survey. Panel attrition could be the result of non-traceable or refusal cases, the impact of which can be quite different from cross-sectional surveys, and these differences should be studied. Allen and Petroni (1994) discuss the problem of adjusting for movers.

There is a need to develop quality studies that take into account the special features of longitudinal surveys. Many quality control studies are available in the conduct of longitudinal surveys besides the usual ones for cross-sectional surveys, since the repeated nature of the study can lead to a more efficient identification of error-prone cases. Since for longitudinal studies, the stability of the data over time is an issue, methodological changes in the study could have an impact on the longitudinal measures that are of interest and these should be evaluated. Administrative data can provide useful evaluations since some of the data can help validate some of the results.

QUESTION 4: ANALYSIS ISSUES

Analysis concerns are the last set of issues we discuss. It is the potential analysis of the longitudinal study that is its most important facet. The causes or determinants of various outcomes are of major interest to the data users. However, the modelling of these causes can be complex, particularly if the survey itself is of a complex nature. Many of these issues are discussed in Singh and Whitridge (1990) and in Hidioglou and Michaud (1998).

Examples of the kinds of analyses that are common would be measures of gross flows or other measures of gross change. Gross flows refers to the change of an individual from one category to another. In other words, it is the flow from category A to category B between two points in time, as opposed to net flow that is the change in the margins over time. There are difficult questions about the impact of measurement error on the measurement of gross flows. If fairly large measurement errors are present on each occasion, there will be a significant impact on the bias of the estimates of the gross flows, even if the net flows themselves are not as adversely affected. Sometimes, sample rotation will aggravate this problem, since accounting for sample rotation properly when measuring gross flows can be problematic. Special treatment is needed for those panels that are entering the sample on a given occasion and for those panels that have left the sample on the previous occasion to get good estimates of these flows. The changes to the population when gross flows are being measured need to be sorted out from the gross flows themselves. In other words, the change from one occasion to another is a combination of the changes in size of the population and the individual changes within the population. The situation can become even more complex when the gross flows are themselves analysed with respect to other information such as income dynamics.

As a referee pointed out, an important issue is the need for educating users on how longitudinal data can be analysed effectively. The recent increase in the number of longitudinal surveys raises many opportunities for new types of analysis, but many analysts who have been studying only cross-sectional surveys may not be aware of the most appropriate techniques.

For the many surveys that use frames based on administrative data, accounting for the frame changes in the analysis may be necessary, since inclusion on the frame can be subject to changes in administrative procedures, as well as changing conditions for the individuals. For example a file of unemployment insurance beneficiaries would be subject to changing eligibility criteria, as well as changing personal situations.

The measurement of change can often be decomposed into various components. For example, the movement of units in the sample from one domain to another can be sorted out from the changes of the data for units within the same domain. Holt and Skinner (1989) contains an interesting discussion on various components of change.

For more complex analyses, such as modelling of time series, most classical time series models do not account for the fact that the information is derived from a sample survey. Therefore, the sampling errors resulting from the sample survey are not properly taken into account in the time series modelling.

In the analysis, some measures may depend on other cross-sectional surveys. For example, it may be another cross-sectional survey that determines the income class

boundaries to be used in the analysis of the longitudinal survey. This may add to the complexity of the analysis since the boundaries can change over time.

Whether and how to use the sampling weights have created difficulties for many analysts, since many of the classical models for analysis of data over time do not use the sampling weights. Procedures need to be developed that incorporate the survey weights in the analysis properly. For large-scale surveys, using the weights is often preferable as this provides some protection against model misspecification.

Errors resulting from the processing, such as the record linkage operation, may need to be incorporated in the analysis or at least some studies need to be taken to understand the impact of these kinds of errors; see, for example Dorinski and Huang (1994).

Often administrative data are used as part of the analysis since these data may be available more readily than collected information. However, since there may be conceptual or other difficulties with the administrative data, special analytical methods may need to be developed to use the administrative data effectively.

Finally, we mention the difficulties associated with the data dissemination. Longitudinal summary measures need to be developed for many phenomena. Often these are not suitable for the usual tabular displays that are commonly used in cross-sectional studies. Many analyses require access to the microdata. This could create problems with respect to protecting the confidentiality of the respondents. The usual measures that one takes when releasing microdata files on cross-sectional surveys may not be sufficient when releasing surveys which are longitudinal in nature, because the databases are so much richer so that the risk of being able to identify an individual on such databases becomes much greater. Protecting the respondents' confidentiality is of paramount importance, so a conservative approach that may not fulfill all the users' requirements may be necessary.

SUMMARY

We have briefly discussed many of the questions and issues that are now being investigated by researchers concerned with the design and analysis of longitudinal studies. Based on our discussion, we see that many questions need to be further investigated. As we gain more experience with longitudinal surveys, many of these issues will be better understood and many new issues will arise. The opportunities for important research and investigation are numerous.

ACKNOWLEDGEMENTS

The author is grateful for many useful suggestions from the referees and the Associate Editor.

REFERENCES

- ALLEN, T.M., and PETRONI, R.J. (1994). Mover nonresponse adjustment research for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 662-667.
- AN, A.B., BREIDT, F.J., and FULLER, W.A. (1994). Regression weighting methods for SIPP data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 434-439.
- ARMSTRONG, J., DARCOVICH, N., and LAVALLÉE, P. (Eds.) (1993). *Proceedings: Symposium 92, Design and Analysis of Longitudinal Surveys*, Statistics Canada.
- BASSI, F., TORELLI, N., and TRIVELLATO, U. (1998). Data and modeling strategies in estimating labour force gross flows affected by classification errors. *Survey Methodology*, 24, 109-122.
- BINDER, D.A., and HIDIROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics Volume 6: Sampling* (Eds. P.R. Krishnaiah and C.R. Rao). North Holland: Amsterdam, 187-211.
- CLARKE, P.S., and CHAMBERS, R.L. (1998). Estimating labour force gross flows from subject to household-level nonignorable nonresponse. *Survey Methodology*, 24, 123-129.
- CORDER, L.S., MANTON, K.G., and WOODBURY, M. (1994). Improving coverage, response rates, and nonresponse follow-up via a longitudinal list sample design: The National Long-Term Care Surveys. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 63-84.
- CZAJKA, J.L. (1994). Income stratification in panel surveys: Issues in design and estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 791-796.
- DORINSKI, S.M., and HUANG, H. (1994). Use of administrative data in SIPP longitudinal estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 656-661.
- DUNCAN, G.J., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.
- FOLSOM, R.E., and WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 428-433.
- HIDIROGLOU, M.A., and MICHAUD, S. (Eds.) (1998). *Proceedings: Symposium 98, Longitudinal Analysis for Complex Surveys*, Statistics Canada. To appear.
- HILL, D.H. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics*, 10, 359-380.
- HOLT, D., and SKINNER, C.J. (1989). Components of change in repeated surveys. *International Statistical Review*, 57, 1-18.
- HUGGINS, V.J., and FISCHER, D.P. (1994). The redesign of the survey of income and program participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.
- HUGHES, S., and HINKINS, S. (1995). Creation of panel data from cross-sectional surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-413.

- KALTON, G., and BRICK, J.M. (1995). Weighting schemes for household panel surveys. *Survey Methodology*, 21, 33-44.
- KALTON, G., and CITRO, C. (1993). Panel surveys: Adding the fourth dimension. *Survey Methodology*, 19, 205-215.
- KASPRZYK, D., DUNCAN, G., KALTON, G., and SINGH, M.P. (Eds.) (1989). *Panel Surveys*. Wiley: New York.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- LENGACHER, J.E., SULLIVAN, C.M., COUPER, M.P., and GROVES, R.M. (1995). Once reluctant, always a reluctant? Effects of differential incentives on later survey participation in a longitudinal study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1029-1034.
- MCUGIGAN, K.A., ELLICKSON, P.L., HAYS, R.D., and BELL, R.M. (1995). Tracking, weighting, and sample selection modeling to correct for attrition. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 402-407.
- MICHAUD, S., DOLSON, D., ADAMS, D., and RENAUD, M. (1995). Combining administrative and survey data to reduce respondent burden in longitudinal surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 11-20.
- MURRAY, T.S., MICHAUD, S., EGAN, M., and LEMAÎTRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference, U.S. Bureau of the Census*, 715-730.
- RIZZO, L., KALTON, G., and BRICK, M. (1994). Adjusting for panel nonresponse in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 422-427.
- SCHEJBAL, J.A., and LAVRAKAS, P.J. (1995). Panel attrition in a dual-frame local area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1035-1039.
- SCHEUREN, F., and WINKLER, W.E. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.
- SINGH, A.C., and WHITRIDGE, P. (Eds.) (1990). *Analysis of data in time. Proceedings: Symposium 89, Analyses of Data in Time*, Statistics Canada.
- SINGH, A.C., WU, S., and BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.
- SINGH, R.P., PETRONI, R.J., and ALLEN, T.M. (1994). Over-sampling in panel surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 674-679.
- STEARNS, S.C., KOVAR, M.G., HAYES, K., and KOCH, G.G. (1996). Estimates of national hospital use from administrative data and personal interviews. *Journal of Official Statistics*, 12, 47-61.
- TAMBAY, J.L., ŞCHIOPU-KRATINA, I., MAYDA, J., STUKEL, D., and NADON, S. (1998). Treatment of nonresponse in cycle two of the National Population Health Survey. *Survey Methodology*, 24, 147-156.
- TIN, J. (1996). Program participation and attrition: The empirical evidence. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 669-674.
- WEBBER, M. (1994). The survey of labour and income dynamics: lessons learned in testing. *Proceedings of the 1994 Annual Research Conference, U.S. Bureau of the Census*, 85-99.

Data and Modelling Strategies in Estimating Labour Force Gross Flows Affected by Classification Errors

FRANCESCA BASSI, NICOLA TORELLI and UGO TRIVELLATO¹

ABSTRACT

Gross flows among labour force states are of great importance in understanding labour market dynamics. Observed flows are typically subject to classification errors, which may induce serious bias. In this paper, some of the most common strategies, used to collect longitudinal information about labour force condition are reviewed, jointly with the modelling approaches developed to correct gross flows, when affected by classification errors. A general framework for estimating gross flows is outlined. Examples are given of different model specifications, applied to data collected with different strategies. Specifically, two cases are considered, *i.e.*, gross flows from (i) the U.S. Survey of Income and Program Participation and (ii) the French Labour Force Survey, a yearly survey collecting retrospective monthly information.

KEY WORDS: Correlated classification errors; Latent class models; Longitudinal data; Recall errors; Seam effect.

1. INTRODUCTION

Gross flows among labour force states, are a powerful tool to analyse labour market dynamics. Gross flows regard changes at individual level, and therefore their estimation rests on the availability of longitudinal data.

The effects of erroneous classification of units with respect to their position in the labour market, can cause spurious transitions. Even if one might assume that these errors cancel out when estimating net flows, they cannot be ignored when estimating gross flows.

Various strategies can be adopted, in order to correct gross flows for classification errors. Basically, they depend on:

- (a) assumptions about the classification error mechanism, following from
 - (a1) the survey design (panel surveys – possibly with a rotating scheme, retrospective surveys, some mixture of retrospective and panel surveys, *etc.*), and/or;
 - (a2) the content and structure of the questionnaire (availability of one or more indicators of the variable of interest, format of the questions – episode based or event based, *etc.*);
- (b) assumptions about the generating process of the transitions among labour force states.

In this paper, some of the most common strategies used to collect longitudinal information about labour force condition are reviewed, jointly with modelling approaches developed to correct gross flows when affected by classification errors. It is shown that most of the usual specifications proposed in the literature, can be seen as special cases of a general formulation, which allows to elucidate advantages and disadvantages of each specification, and makes it possible to consider a common estimation strategy.

The focus of the paper is on sound applications of this general modelling approach, for estimating gross flows from survey data collected with different strategies. Two cases are considered: (i) the U.S. Survey of Income and Program Participation and (ii) the French Labour Force Survey, a yearly rotating panel survey with retrospective monthly information.

The organization of the paper is as follows. Section 2 briefly discusses various strategies for collecting longitudinal data on labour force participation, and their likely implications for classification errors, as they emerge from the survey methodology literature. In section 3, a fairly general approach for modelling gross flows affected by classification errors, *i.e.*, for jointly estimating true gross flows and conditional response probabilities, is outlined. Examples are also given on how some well known models for correcting observed gross flows, can be specified as special cases of this approach (section 3.1). Attention is then devoted to a convenient framework for formulating the above models, provided by latent class models and, more specifically, by the so-called “modified LISREL model” proposed by Hagenaars (1990), a general tool to describe causal relationships among observed and unobserved categorical variables (section 3.2).

The final, and main part of the paper (section 4), is devoted to a detailed presentation of the two case-studies. The modelling approach is common: *a priori* information on the measurement characteristics of the survey (and possibly on the true process), is combined with specification searches, in order to obtain parsimonious and (hopefully) sensible models. As already noted, the two case-studies are reasonably different, chiefly in terms of the design of the surveys: this diversification turns out to be useful for illustrating different model specifications, and various strategies for reaching/testing the final formulation.

¹ Francesca Bassi, Nicola Torelli and Ugo Trivellato, Dipartimento di Scienze Statistiche, Via San Francesco, 33, 35121, Padova, Italy.

From the two case-studies, the following overall evidence can be drawn:

- (a) the modified LISREL model has proved to be a set-up, flexible enough for modelling the error mechanism in longitudinal data collected with different survey designs, as well as the generating process of true labour force transitions;
- (b) specifically, in the measurement part of the model, we were able to incorporate the pattern and the effects of correlated classification errors, which are particularly important in surveys with retrospective features;
- (c) observed transitions are corrected towards the direction expected, on the basis of theoretical and empirical evidence on measurement errors effects, (not mechanically towards mobility, as strategies based on the assumption of independent classification errors do).

2. THE ROLE OF DATA COLLECTION STRATEGIES

Information for labour gross flows estimation comes from longitudinal data, *i.e.*, observations on the same units pertaining to different time points. Recently, there have been increasing efforts in collecting longitudinal data. This is true also for surveys, whose main goal is to measure the labour force condition of individuals in a given population. On the other side, this focus on collecting, and using longitudinal data, raised new questions about the origin and pattern of measurement (= classification) errors, as well as their possible effects on estimates of the quantities of interest. General references about sources of classification errors for longitudinal data, collected by surveys across time, are Duncan and Kalton (1987) and Kalton and Citro (1993). In this section, some main implications of classification errors on modelling strategies, to correct gross flows are briefly discussed.

A typical argument about the effect of measurement error in estimating gross flows, is that it leads to over-estimation of changes. This is true when one assumes that measurement errors are not correlated over time. This assumption is not realistic in many cases (see Skinner and Torelli 1993; Singh and Rao 1995; van de Pol and Langeheine 1997), and should be reconsidered taking carefully into account, the data collection strategy actually adopted. Broadly speaking, if longitudinal data are (at least partly) collected by retrospective interrogation, one can argue that memory inaccuracy leads to correlated errors.

Specific assumptions about classification errors can be successfully introduced in appropriate statistical models, only if additional information is available in the form of plausible *a priori* knowledge about the error generating mechanism and/or supplementary data about the labour force state.

Modelling strategies to correct gross flows for classification errors, should then take into account the measurement process actually used, in the sense that the amount of classification errors and the direction of possible bias, are related to the strategy adopted to collect longitudinal data.

As it is well known, longitudinal data can be obtained by different survey strategies. It is convenient to distinguish at least between (i) panel surveys and (ii) retrospective surveys. In addition, the availability of multiple indicators deserves specific attention.

Panel surveys are the most natural ways of collecting longitudinal information. Among these, rotating panel surveys play a prominent role. In fact, this is the scheme adopted in most national Labour Force Surveys (LFSs), whose primary goal is estimation of labour force stocks. For LFSs with a rotating sampling design, longitudinal information on the (usually short) sequence of states, can be easily obtained by matching data on individuals participating in two or more successive surveys. In LFSs, the reference period, concepts and definitions for classifying people, are typically consistent with the International Labour Office (ILO) recommendations (Husmanns, Mehran and Verma 1990): this makes measures of labour force conditions reasonably accurate and comparable over space and time. Data on labour force participation are collected also through general purpose household surveys. In this case, attention to labour force condition is less prominent than in the preceding type of surveys, and reference periods, concepts and definitions, might be less consistent with ILO recommendations.

Alternatively, longitudinal information can be collected by retrospective surveys. Cross-sectional surveys can include retrospective questions, to get information on the sequence of labour force states experienced by sampled individuals. In this case, the interrogation strategy is crucial to reduce errors due to memory (recall errors, telescoping, *etc.*). Procedures to improve accurate reporting in retrospective surveys, rely upon contributions from cognitive psychology and survey methodology (for a review, see O'Muircheartaigh 1996). Besides, evidence on the amount and the direction of bias due to memory inaccuracy, is found in many empirical studies. It is worth adding, that in retrospective surveys, factors related to length of recall period, salience of events considered, and/or difficulty in retrieving data on past events, usually lead to a simplified format of questions, not consistent with ILO conventions on labour force condition.

Interesting opportunities for estimating gross labour flows in the presence of classification errors, come from the widespread practise of using a mixture of the panel and the retrospective strategies. Panel surveys use retrospective questions, at least on a limited number of topics, to cover the period between two successive waves (this is the case of the Survey of Income and Program Participation, as will be seen in section 4.2). The main characteristics of the measurement process when such a mixed strategy is used,

have to be carefully considered, as they might have a considerable impact in formulating reasonable models for classification errors. More specific traits of the measurement process emerge also from consideration of the peculiarities of the survey design.

From a different perspective, an important opportunity for modelling classification errors is given by the availability of multiple measurements of labour force state, *i.e.*, data on the labour market condition of an individual at a given time, provided by two or more different sources. This information is of great importance in general, and particularly when fairly complicated patterns of correlated classification errors are to be considered. Multiple indicators on labour force state can be collected (i) in the same interview or (ii) in different interviews (*e.g.*, in different waves of a panel survey).

The first case is not very common, but sometimes questions regarding labour force condition are asked in different contexts, and in different ways. For instance, first, a self-classification of the individual with respect to labour force condition is asked; then, in a different section of the questionnaire, a sequence of questions are put forward that allow to classify the respondent according to standard labour force definitions. (For a different example, see the case of the Survey of Income and Program Participation in section 4.2.)

The second case covers several situations. At least two of them are worth considering:

- (a) data from reinterview studies, often collected specifically to get information on classification errors probabilities (in such a case, the common practice is to assimilate reinterview data to validation data: for classical procedures to correct gross flows based on reinterview data, see Abowd and Zellner 1985, Poterba and Summers 1986, and Chua and Fuller 1987);
- (b) data collected retrospectively in panel surveys, but referring to a time point already covered by the preceding interview, or collected in a supplementary survey carried out occasionally and covering the reference period(s) of the current panel survey. It is obvious that, in this case different measures of the same variable(s) of interest can be polluted by classification errors with largely different characteristics.

Many of the points raised here will be clarified in the case-studies presented in section 4, where the joint presence of panel and retrospective information and of multiple indicators of the same latent variable is exploited in order to get parsimonious models.

3. ESTIMATING GROSS FLOWS AFFECTED BY CLASSIFICATION ERRORS

3.1 A General Framework

Specification of statistical models to adjust labour force gross flows for classification errors, should allow one to take into account, the nature of available data (as reviewed in the previous section), and substantial assumptions on the generating process of (i) transitions among labour force states (*e.g.*, Markov chain structures) and (ii) measurement errors (*e.g.*, uncorrelated *vs.* correlated measurement errors).

In the simplest case, we consider panel data, where at each time period $t = 1, \dots, T$, a discrete variable Y_t is observed for a generic unit, in a random sample of size n . In our case-studies, the units will be individuals, and the time periods, months or quarters. Y_t takes one among r possible distinct values or states. Y_t is an imperfect measure of y_t , which denotes the true state of a generic unit at time t . In general, it is not necessary to assume, that y_t varies over the same set of states $1, 2, \dots, r$, but for simplicity, and without loss of generality, we will consider here the same set of states as for Y_t .

Strategies for estimating gross flows, rely upon an appropriate specification of the joint probability of the true and the observed process $P(Y_1, \dots, Y_T, y_1, \dots, y_T)$. Statistical analysis is then based on marginalization with respect to unobserved quantities:

$$P(Y_1, \dots, Y_T) = \sum_{y_1=1}^r \dots \sum_{y_T=1}^r P(Y_1, \dots, Y_T, y_1, \dots, y_T). \quad (3.1)$$

Models are based on parsimonious specifications of the joint probability function $P(Y_1, \dots, Y_T, y_1, \dots, y_T)$. Essentially this can be obtained by decomposing it into a product of conditional probabilities, following from an appropriate set of assumptions about the dependence structure among the components $Y_1, \dots, Y_T, y_1, \dots, y_T$.

For our purposes, a convenient starting point for model specification, comes from assumptions (i) about the structure of the generating process of the true transitions among labour force states and (ii) about the measurement process (exploiting, for instance, substantial knowledge or empirical evidence from the data collection strategy adopted).

In a model aimed at distinguishing between true and observed turnover in the labour market, a typical example that exploits this idea, is provided by Latent Class Markov (LCM) models (van de Pol and Langeheine 1990). For a generic unit, the following probabilities are specified:

$$q_t^{l_t j_t} = P(Y_t = l_t | y_t = j_t) \quad t = 1, \dots, T \quad (3.2)$$

$$\pi_t^{j_t, j_{t-1}} = P(y_t = j_t | y_{t-1} = j_{t-1}) \quad t = 2, \dots, T \quad (3.3)$$

$$\pi_1^{j_1} = P(y_1 = j_1) \quad (3.4)$$

Conditional probabilities (3.2) represent the relationship between true and observed states, *i.e.*, the probability of reporting at time t , state l_t , while the true state is j_t . Clearly, this specification implies the local independence assumption, *i.e.*, Y_1, \dots, Y_T are independent, given y_1, \dots, y_T . Conditional probabilities (3.3) describe the dynamics in the labour market, *i.e.*, the probability that a transition from j_{t-1} to j_t occurs, when moving from time $t-1$ to t : according to (3.3), the true transition process evolves following a first order Markov chain. Finally, probabilities (3.4) describe the initial condition for the Markov process.

The marginal probability for the observed sequence (3.1) is then given by:

$$P(Y_1 = l_1, \dots, Y_T = l_T) = \sum_{j_1=1}^r \dots \sum_{j_r=1}^r \pi_1^{j_1} \prod_{t=2}^T q_t^{l_t, j_t} \pi_t^{j_t, j_{t-1}} \quad (3.5)$$

For four measurement points, model (3.5) is equivalently represented by the path diagram in Figure 1, where arrows indicate direct effects between variables.

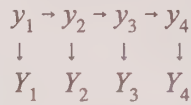


Figure 1. Path Diagram of a LCM Model for Four Measurement Points

It is worth observing, that the assumption of local independence is equivalent to the Independent Classification Errors (ICE) assumption. As noted in the previous section, the ICE assumption has been severely criticised, and seems definitely unreasonable when longitudinal data are collected by retrospective questions.

As another example, for $T = 2$, classical strategies to correct gross flows based on reinterview studies, can be represented within the framework outlined above. In this case, additional information is used, in the sense that the q_t parameters are exogenously estimated from the reinterview study, and are plugged in (3.5) in order to obtain directly $P(y_1, y_2)$.

The same framework can be used, to encompass more general assumptions on both the latent and measurement processes, up to include serially correlated classification errors. As an interesting case, we consider the model by Pfeffermann, Skinner and Humphreys (1998). Ignoring here initial conditions, they reformulate conditional response probabilities as follows:

$$q_t^{l_t, j_t} = P(Y_t = l_t | y_t = j_t, Y_{t-1} = l_{t-1}) \quad t = 2, \dots, T, \quad (3.6)$$

thus overcoming the ICE assumption.

A similar formulation, aimed at introducing, at least partially, dependence between the observed state at time t and the sequence of true states at times t and $t-1$, has been suggested by van de Pol and Langeheine (1992), who extend the model to allow also for a second order Markov chain, for the true transition process.

The modelling strategy for estimating true flows can be further extended in various directions, namely:

- It is straightforward to extend the model, to exploit the availability of multiple indicators of the same unobserved true state. This implies that response probabilities, as those in (3.2), are defined for one or more additional observed variables, treated as imperfect measures of the same latent state y_t . As an example, a LCM model for two indicators per latent variable, and four points in time, is represented in Figure 2. In this model, each couple of indicators referring to a given point in time, is assumed to be independent, conditionally on the corresponding latent variable, in the sense that the correlation between them, is completely explained by their relation with y_t .
- Observed heterogeneity at the individual level, in the transition and/or the measurement processes, can be introduced by conditioning on a set of covariates X_t . An example is given in Pfeffermann *et al.* (1998). They use covariate information at the unit level and model their impact on labour market condition by multinomial logit.
- Unobserved heterogeneity can also be considered, which leads to mixed latent class models (van de Pol and Langeheine 1990). A simple case is the movers/stayers model, where a different behaviour, at the latent level, is assumed for groups of units, while the group membership of the units cannot be directly observed.

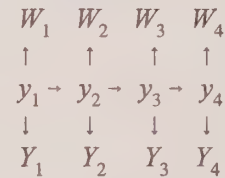


Figure 2. Path Diagram of a LCM Model for Four Measurement Points and Two Indicators for Each Latent Variable

3.2 Latent Class and Related Models as a Tool for Estimating Gross Flows With Measurement Errors

A special case of the general model formulation outlined in the above section, are latent class models, where the true state in the labour market plays the role of the latent variable, and the observed state acts as its indicator. Some of the specifications outlined in the previous section, include dependence among classification errors. A general and convenient approach for handling it, which includes standard latent class models with correlated classification

errors, is the so called modified LISREL model proposed by Hagenaars (1990).

The modified LISREL approach consists of an extension of Goodman's (1973) path analysis, which is a tool to describe causal relationships among observed categorical variables, through a system of logit equations. Basically, the extension incorporates latent variables. Thus, a modified LISREL model combines a measurement sub-model, which specifies the dependence of the indicators on latent variables, and a structural sub-model, which specifies ordered relations among latent and possible external variables. As the name itself suggests, it can also be viewed as the analogue for discrete variables, of the well known LISREL model for continuous variables (Joreskog and Sörbom 1988).

Modified LISREL models, allow to introduce serially correlated classification errors, by inserting direct effects between the indicators (Hagenaars 1988). The presence of direct effects implies, that the association among observed variables, is not completely explained by the effects of the latent variables on their indicators, but that there exists a source of additional association among the indicators, over and above the part that is explained by their relation with the latent variables.

Once a reasonable model has been specified, identification should be ascertained. The model involves many unobservables, and identification of all parameters is not automatically assured.

Reasonable opportunities to achieve identification, rest on two strategies, possibly used in combination: (i) imposition of plausible equality restrictions among the set of parameters and (ii) availability of multiple indicators of the unobserved true state. The latent class Markov model represented in Figure 1, for example, is not identified without extra restrictions on its parameters. If the latent chain is assumed to be time homogeneous, or response probabilities are restricted to be equal across time, the model can be shown to be identified (Lazarsfeld and Henry 1968). Availability of multiple indicators for the unobserved true state, can also help identification of complex measurement models. Identification criteria for some very special specifications, have been proven (for example, the model in Figure 2 can be shown to be identified), but no general rules have been provided yet to ascertain global identification. It is advisable to check at least local identification, *i.e.*, identifiability of the unknown parameters in a neighbourhood of the maximum likelihood solution. Goodman (1974) stated that a sufficient condition for local identifiability of a latent class model, is that the Information matrix be full of rank. Goodman's condition may be computationally difficult to check. Moreover, with some data sets, it may happen that the Information matrix is not of full rank, simply because some estimates are very close to the boundaries of the parameter space. An alternative, empirical way to check identifiability, is to estimate the model using different sets of starting values. If

different sets of starting values result in the same value for the log-likelihood function but in different parameter estimates, then the model is not identifiable.

As for estimation, modified LISREL models may be treated as directed loglinear models with latent variables (Hagenaars 1997). A directed loglinear model results in a sequence of parsimonious multinomial logit models, possibly with latent variables, which are estimated stepwise. At each step, one dependent variable is considered, and a multinomial logit model is estimated on a contingency table, which has been collapsed over the variables, that do not directly influence the dependent variable in the causal order. Estimates obtained at each step are, at the end, combined in order to obtain estimated parameters for the full model. Directed loglinear modelling yields exactly the same parameter estimates, standard errors and test statistics as the Goodman standard procedure, but using simpler marginal tables. If the causal model contains one or more latent variables, an appropriate estimation technique must be used, *e.g.*, an implementation of the EM algorithm (Meng and Rubin 1993).

The empirical validity of the complete causal model may be tested, comparing the estimated expected frequencies with the observed ones in the complete table, by means of the likelihood ratio L^2 and the Pearson X^2 statistics. However, the structure of the observed data on labour market transitions, is such that many cells show very low observed frequencies. For this reason, the usual X^2 and L^2 criteria must be used only as a general indication of fit, since their asymptotic χ^2 distribution is no longer guaranteed, due to the sparse and unbalanced pattern of the contingency table.

Various strategies can be adopted to extend and improve model evaluation, and three of them are worth mentioning in this context:

- (i) A restricted model nested within a larger one, can be tested with the conditional test, *i.e.*, considering the difference in the L^2 values of the two models, which is asymptotically distributed as χ^2 under weaker conditions (Goodman 1981, and Haberman 1978).
- (ii) In general, using multiple criteria can be a sensible strategy. Indices based on the information criterion, such as AIC or BIC, can be useful to compare alternative non-nested models. Another advantage of AIC and BIC is that, in the selection procedure, they weight the goodness of fit of a model against its parsimony, considering the model degrees of freedom and the sample size. (AIC = $L^2 - 2 \times$ degrees of freedom. BIC = $L^2 - \ln(N+1) \times$ degrees of freedom.) The model that is preferred, in this context, is the one with the lowest value of AIC or BIC.
- (iii) Monte Carlo resampling techniques can be implemented to simulate the asymptotic distribution of X^2 and L^2 (Langeheine, Pannekoek and van de Pol 1995).

4. TWO CASE-STUDIES

4.1 The General Set Up

In this section we present two applications of the modified LISREL approach to correct observed gross flows in the labour market. Data come from surveys with partly different designs:

- (1) the U.S. Survey of Income and Program Participation (SIPP), a multi panel household survey, which collects retrospective information on the between waves working history;
- (2) the French Labour Force Survey (FLFS), a yearly retrospective survey, with one month overlapping reference periods.

For each case-study, a model is specified on the basis of *a priori* information on both the true transition process and the error generating mechanism. *A priori* information is crucial for model specification, in order to obtain parsimonious and plausible models.

All the models are written in the form of a modified LISREL model, and estimated by the EM algorithm. Actually, we used the IEM program (Vermunt 1993) and checked all the models for local maxima.

The two final models turn out to be rather complex, since they incorporate correlation among classification errors, and specific assumptions on respondent's behaviour. This fact, together with the sparse and unbalanced pattern of the observed contingency table, typical of labour force transitions, demands for goodness of fit evaluation criteria, other than L^2 and X^2 . In the first case-study, alternative models have been judged by means of the BIC index, and on the basis on substantive knowledge on the labour market in the U.S.. In the second case, alternative models have been compared by means of the conditional test.

In the following sections, models are presented in a logical and verbal form, while the mathematical formulation for the final model is given in the relevant Appendix.

4.2 The SIPP Data

SIPP is a multi panel household survey conducted by the U.S. Bureau of the Census, in order to collect information on topics such as employment, income, participation in social programs, *etc.* The reference population is the U.S. noninstitutionalized individuals over 14.

The survey started in 1984, and is a continuing one: as a general pattern, each year a new sample of households, called "panel", has been selected for the survey and followed for two and half years (for a detailed description of SIPP, see U.S. Department of Commerce 1991, and Citro and Kalton (1993)).

Each panel is randomly divided into four "rotation groups" and interviewed at 4-months intervals for eight times. For practical reasons, each rotation group is interviewed in each of four consecutive months, and retrospective questions collect information with reference to the 4-months period elapsing between subsequent interviews. Each set of interviews with the full sample is termed a "wave".

We will refer to the 1986 panel, which started in February 1986 and ended in August 1988. We will consider the intermediate period from January 1986 to January 1987, over which we have information from all four rotation groups. Figure 3 represents the survey design with regard to our sample.

Information on labour force participation, is collected mainly in the "Labour Force and Reciprocity" section of the questionnaire (for an additional piece of information, collected in another section of the questionnaire, see below), where each respondent is asked to report on a weekly basis his/her labour market history in the preceding four months (18 weeks), by going through a series of filtered questions. The respondent is first asked whether he/she had a job or a business, at any point in time during the reference period. If the respondent gives a negative answer, he/she is asked whether he/she spent any time looking for work, or was in layoff, and, if so, in exactly which weeks. On the other hand, if the answer to the

Interview Month	Rot. Group	Wave	Reference months							
			Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
February	2	1								
March	3	1								
April	4	1								
May	1	1								
June	2	2								
July	3	2								
August	4	2								
September	1	2								

Figure 3. Rotation Plan for the 1986 SIPP Panel (First 2 Waves)

starting question is positive (*i.e.*, he/she worked some time), and the respondent declared a job or a business with continuity during the reference period, he/she will move to the following section of the questionnaire. The respondent not declaring a stable situation in the labour market, is asked a long series of questions in order to establish the labour force state occupied, in each single week of the reference period.

The weekly based information is usually recorded, to obtain a monthly classification based on the usual three categories: Employed (E), Unemployed (U) and Not in the labour force (N). For individuals covering different positions during one month, the monthly labour force state is the one identified by the “modal” category with regard to the weeks of that month (Martini 1989).

Observed gross flows between two generic calendar months are then obtained as follows:

- (a) For individuals belonging to three rotation groups, on the basis of retrospective data collected in the same interview. These observed flows will be called “within wave” (WW) transitions.
- (b) For individuals in the fourth rotation group, by combining information collected in two different interviews, four months apart. These observed flows are termed “between waves” (BW) transitions.

When estimating monthly changes, a peculiar problem with SIPP data, is the so called “seam effect” (Young 1989): more changes are observed when data for two adjacent months are collected in two different waves – the transition covers the seam of the waves – than when they come from the same interview. The seam effect is pervasive in the survey: evidence of it for several variables of interest, is reported in Martini (1988), Marquis and Moore (1989), Kalton and Miller (1991).

Table 1 illustrates this phenomenon for our 1986 SIPP panel sample. Row 4-1 contains average BW transition rates; rows 1-2, 2-3 and 3-4 contain average WW transition rates, pertaining to the position of the two relevant reference months in each wave (for example, row 1-2 contains transition rates between the first two reference months in each wave). From Table 1, there is clear evidence that observed WW transitions describe a more stable labour market than BW ones. Moreover, WW stability increases, moving backwards in the wave (from 3-4 to 1-2).

One reasonable explanation for the seam effect, and for the systematic pattern of observed transitions throughout a wave, is the different role of measurement errors, for data obtained under the BW and WW strategies respectively. Specifically, it is likely that classification errors have a different degree of correlation for WW and BW observed flows: the higher stability documented by WW transitions may be induced by highly correlated classification errors. Indeed, if errors were uncorrelated, specifically for WW transitions, no evidence of seam effect would be expected.

A variety of plausible causes of correlated errors, is suggested by the cognitive psychology and the survey methodology literature on memory effect and recall errors (see, Bernard, Killworth, Kronenfeld and Sailer 1984, and O’Muircheartaigh 1996), among which a “conditioning” effect: respondents tend to give the same answer going backwards within the wave, and in extreme cases, they mechanically repeat the same answer for all four months.

Table 1

Observed Monthly Transition Rates ($\times 100$) for the 1986 SIPP Panel, January 1986 to January 1987

	Type	EE	EU	EN	UE	UU	UN	NE	NU	NN
1-2	WW	98.27	1.04	0.69	15.46	79.63	4.91	1.15	1.42	97.43
2-3	WW	97.91	1.13	0.96	17.34	75.96	6.70	1.38	1.71	96.91
3-4	WW	97.85	1.20	0.95	19.23	73.25	7.52	1.28	1.69	97.03
4-1	BW	94.03	2.10	3.87	26.81	42.20	30.99	5.65	3.77	90.58

Abundant empirical literature shows, that this sort of conditioning effect is the main source of classification errors in SIPP data. Other potential sources of error, typical of panel surveys, do not affect SIPP data dramatically. Administrative record check studies find little, if any, evidence of time-in-sample effect (Chakrabarty and Williams 1989; McCormick, Butler and Singh 1992). As a general consideration, we may say that in SIPP data, the seam effect dominates over other sources of error, that potentially bias gross flows estimates.

Summing up, a model-based approach to obtain unbiased gross flows from SIPP data, is justified by two arguments:

- (a) the patent presence of correlated classification errors;
- (b) *a priori* information on the data generating mechanism, drawn from two sources:
 - (b1) specific evidences emerging from SIPP observed gross flows, such as the seam effect, and the increase in stability going backwards within the wave, just documented;
 - (b2) general hints provided by the social survey literature on respondent behaviour.

In order to correct SIPP observed labour force gross flows from classification errors, a model has been built, based on the following assumptions/information:

- (a) the true transition process follows a first order Markov chain;
- (b) WW data transitions are affected by correlated classification errors, according to a pattern that will be specified in the sequel;
- (c) for BW, the standard ICE assumption holds;
- (d) rotation groups are equivalent samples also for modelling purposes, *i.e.*, respondents behave in the same way in all four rotation groups;
- (e) SIPP data provide two indications on the monthly labour force state of each individual: the detailed information collected in the “Labour Force and Reciprocity” section

of the questionnaire, just presented, and the additional information collected in the “Earnings and Employment” section, where the respondent is asked if he/she did/did not have a job in the reference period, on a weekly basis.

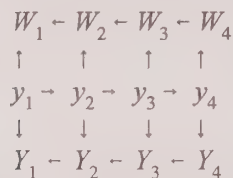


Figure 4. Path Diagram of a Modified Lisrel Model for Four Measurement Points and Two Indicators for Each Latent Variable (for the Meaning of Symbols, See Main Text)

Figure 4 contains the path diagram of a simplified version of the model (*i.e.*, a version that does not aim at representing in detail, the pattern of correlated classification errors, nor at taking into account the fact that we are dealing with four rotation groups) for four points in time, *i.e.*, for four consecutive calendar months. Here y_t ($t = 1, 2, 3, 4$) represents latent variables; Y_t and W_t represent indicators; arrows indicate direct effects between pairs of variables. Indicator Y_t refers to the reported labour force state, described by the usual three categories (E, U and N), while W_t refers to the binary variable Job/No Job. Since information is collected in two different sections of the questionnaire, and with different interviewing procedures, Y_t and W_t can be assumed to be independent given y_t . On the other hand, direct effects between the indicators, account for correlated classification errors over time: the response given for time $t + 1$ affects that given for time t . Note also, that an additional variable G with four categories should be added to the diagram, to account for rotation group membership. All indicators depend on G , since units in different groups are interviewed in different calendar months.

The basic equation of the model, decomposes the proportion in the generic cell of the 9-way contingency table, in the product of the conditional probabilities reported in Appendix A, equations (A1) to (A7). A preliminary version of the model has been proposed in Bassi, Croon, Hagenaars and Vermunt (1995).

Equation (A1) defines the probability of belonging to one of the four rotation groups. Equations (A2) and (A3) define the initial condition, and the transition probabilities, of the latent first order Markov chain respectively. Equations (A4) and (A5) define the response probabilities for indicator Y_t ; equations (A6) and (A7) the analogous probabilities for the dichotomous indicator W_t . The response probabilities are defined in such a way that the answer given for a certain month, depends jointly on the current true state (y_t) and on the “past” true and “past” reported states (y_{t+1} and Y_{t+1}). The term “past” refers to the way respondents think, while answering retrospective questions: they start recalling from the moment of time

nearest to the interview, and go backwards up to the end of the reference period.

A complex set of constraints has been imposed on response probabilities of (A4), (A5), (A6) and (A7), to account for (i) the conditioning effect, and (ii) the fact that the four rotation groups are equivalent samples in terms of the error generating mechanism.

These constraints are formulated in detail in Appendix A. Basically, they incorporate *a priori* knowledge on respondent’s behaviour, and allow us to specify a parsimonious model. Specifically, equations (A8) to (A14) correspond to the following statements:

- (a) With regard to WW classification errors, following Hubble and Judkins (1989), it is assumed that:
 - (a1) a respondent who reports wrongly his/her labour force state for a certain month, continues to repeat this same answer also for the adjacent month, going backwards within the wave (A8);
 - (a2) if, however, the status at time $t + 1$ is correctly reported, the response probability for the adjacent month depends only on the current true state (A9);
 - (a3) the same error generating mechanism operates for both indicators. For W_t , we state that a correct answer is given when the true state is E and ‘Job’ is reported and when the true state is U or N and ‘No Job’ is reported, (A10) and (A11).
- (b) Response probabilities are set equal across rotation groups, (A12) to (A15). As an example, equalities in (A12) mean that response probabilities for individuals in rotation group 1 for the month of April, are equal to response probabilities for individuals in group 4 for the month of March, to those for individuals in group 3 for the month of February, and to those for individuals in group 2 for the month of January. (They are set to be equal, since they all refer to the answer given for the last month of the wave.)

The model has been estimated to correct observed monthly gross flows for the quarter January to April 1986 (Table 2). The comparison between observed and estimated flows, highlights that the model reduces the seam effect: WW transitions are corrected towards a more dynamic labour market; BW transitions are corrected in the opposite direction. It is worth noting, that effects of model correction are more evident for flows from unemployment, which are characterised by higher mobility.

The goodness of fit of the model has been judged by multiple criteria such as the BIC index and the conditional test for nested models, together with estimate interpretability and consistency, with substantive knowledge of the dynamics of the U.S. labour market in ‘80s.

4.3 The French Labour Force Survey Data

The second case-study refers to the flows in the labour market, observed with the French Labour Force Survey (FLFS) conducted yearly by INSEE in France.

Table 2
SIPP Observed and Estimated Monthly Transition Rates
($\times 100$), January to April 1986

		EE	EU	EN	UE	UU	UN	NE	NU	NN
J-F	WW	98.11	1.17	0.72	14.53	80.16	5.31	0.90	1.57	97.53
	BW	94.08	2.17	3.75	23.58	44.30	32.12	5.62	3.45	90.93
	Estimated	97.25	1.47	1.28	16.08	77.16	6.76	1.59	1.32	97.09
F-M	WW	98.66	0.92	0.42	16.06	78.67	5.27	0.64	1.65	97.71
	BW	94.88	1.91	3.21	21.90	48.54	29.56	4.99	4.11	90.90
	Estimated	97.83	1.20	0.97	19.40	74.01	6.59	1.21	1.50	97.29
M-A	WW	98.71	0.64	0.65	20.76	71.74	7.50	1.47	1.05	97.48
	BW	95.59	1.52	2.89	30.48	34.92	34.60	6.34	3.78	89.88
	Estimated	98.11	0.95	0.94	26.42	65.75	7.83	2.17	0.71	97.12

The reference population of the FLFS are all members of French households, who are above 15 in the year in which the interview is planned. The survey has a rotating design: each year, one third of the sample is renewed.

Information on labour force participation is collected with retrospective questions, having as a reference period the 13 months preceding the interview. Each respondent is asked to recall his/her position in the labour market on a monthly basis, by filling in a grid in which he/she can classify himself/herself, for each month, over eight categories: self-employed, employed on a fixed term basis, permanently employed, unemployed, on training, student, serving in the Army, other (retired, housewife, *etc.*).

For our analysis, we aggregated the eight categories in the usual three states E, U and N. We consider 'Employed' respondents who classify themselves in the first three categories, 'Unemployed' those who classify themselves in the fourth category and 'Not in the labour force' the remaining ones.

We analyze the information collected in the two consecutive waves of March 1991 and March 1992, on a subsample of individuals: those who answered to three consecutive interviews (January 1990, March 1991 and March 1992) and who were 18 to 29 years old in 1992, for a total of 5,427 individuals. The reference periods of the two waves considered, overlap in March 1991. We have, then, two pieces of information on the labour force state for this month: one collected in March 1991, and the other one collected with a retrospective question 12 months afterwards.

The pattern of observed monthly transitions in our FLFS sample shows some interesting evidence, largely dictated by the characteristics of the subsample – young people.

Transitions exhibit a moderate degree of seasonal variation, related to the school calendar. From June to July, for example, we observe a proportion of people who enter the labour market as employed, greater than the average; on the contrary, from August to September, a proportion greater than the average leaves employment (presumably to education).

The marginal distribution of the three states from March 1990 to March 1992, shows that the individuals in our sample progressively enter the labour market: in March 1990, 44% are observed to be Employed or Unemployed, whereas by March 1992, this proportion has risen to 54%.

The double information for March 1991, provides some crude evidence on response error in the data: 8% of respondents declare a different state in the two interviews. For the period from February to April 1991, two types of flows may be observed: a within wave (WW) one, *i.e.*, information about the labour force state is collected in the same interview, and a between waves (BW) one, *i.e.*, information is collected in two different interviews (Table 3).

Table 3
FLFS Observed Monthly Transition Rates ($\times 100$) from
February to April 1991

		EE	EU	EN	UE	UU	UN	NE	NU	NN
F-M	WW	98.19	1.67	0.14	9.11	90.65	0.24	0.28	0.11	99.61
	BW	93.17	3.58	3.25	25.18	65.23	9.59	3.75	1.96	94.29
M-A	WW	98.60	1.04	0.36	8.89	90.37	0.74	0.24	0.29	99.47
	BW	93.24	3.33	3.43	25.90	63.79	10.31	3.79	2.07	94.14

As expected, WW transitions describe a more stable labour market than the BW ones. This can be considered as an indication of correlated classification errors in the data. Patterns and causes of errors correlation in retrospective surveys, have been extensively discussed in the two previous sections, and the above considerations can largely be extended to the FLFS data.

In general, we expect that, in a retrospective survey with such a long recall period, lack of memory results in the major cause of classification errors. We also expect that the probability of answering incorrectly, increases as the distance between the reference month and the interview month gets longer. This may be considered as the major source of correlation among classification errors, together with telescoping and conditioning effects, which possibly affect FLFS data as well (see Magnac and Visser 1995).

The overall effect of correlated classification errors, reasonably results in an underestimation of mobility in the French labour market.

Moving from these considerations, we specified a model to correct observed quarterly gross flows, from measurement error (Table 4). The last column of Table 4 contains the percentage of individuals who are observed to change state, between the two months considered (OM = observed mobility). On the average over the five WW transitions, 6.122% of mobility between two consecutive months is observed.

As in the previous case-studies, let us denote with y_t ($t = 1, 2, 3, 4, 5, 6$) true labour force states, and with upper case letters their indicators: Y_t ($t = 2, 3, 4, 5, 6$) represents labour force states observed in March 1992 (referring to March, June, September, December 1991 and March 1992); W_t ($t = 1, 2$) represents labour force states observed in March 1991 (referring to December and March 1991). As usual, y_t , Y_t and W_t distribute over the three categories of E, U and N.

The model is specified by decomposing the proportion in the generic cell of the 7-way contingency table as in Appendix B, equations (B1) to (B6).

Since we observe two indicators only for one month, a model which assumes direct effects between the indicators, would be under identified. Thus, we can not explicitly model dependencies between observed states. The only way to account for correlated classification errors in FLFS data, is to let observed states depend on latent transitions. By the way, this seems to be a sensible assumption in retrospective surveys. Indeed, flows between two different states may easily undergo wrong placements in time, because in some situations, events might truly be difficult to place exactly. As an example, employees who loose their job or retire (flows EU and EN), will generally use the holidays they are entitled to, and may not clearly know when they exactly left employment. The moment people entered the labour force, may also be hard to recall, especially when they left school (flows NU and NE) (van de Pol and Langeheine 1997).

The modified LISREL model, formulated in mathematical terms in Appendix B, is based on the following substantive assumptions.

At the latent level, transitions follow a first order non stationary Markov chain (equations (B1) and (B2)). Indeed, the evidence on seasonality in observed transitions, suggest avoiding the imposition of stationarity of any order, on the latent Markov chain.

Response probabilities for data collected in both waves, depend on the latent transition occurring between t and $t + 1$ (equations (B3) and (B4) refer to data collected in March 1992, equations (B5) and (B6) to data collected in March 1991).

In order to describe the error generating mechanism in detail, and specify a more parsimonious model, the following constraints have been imposed on response probabilities:

- response probabilities referring to the same month of subsequent years (December and March) are set equal;
- response probabilities at time t , given that the true state has not changed between time t and time $t + 1$, are set constant over time;
- response probabilities are set equal for June and September 1991;
- in general, respondents who move between month t and $t + 1$ (transitions EU, EN, UE, UN and NU), at time t , report either the true state occupied at time t , or the true state occupied at time $t + 1$, *i.e.*, they, do not report a state they have not been moved from/to;
- if however, the latent transition occurs between states N and E, we admit all three answers at time t , *i.e.*, we consider that people who find a job may confuse their previous position (at time t), and be uncertain between U and N.

Constraint (c) is imposed mainly for reasons of model parsimony. It captures the notion that response probabilities for months that are placed more or less in the central part of the reference period, do not vary too much.

Constraints (b) and (d) reflect the fact that response probabilities depend on latent transitions. We expect that these probabilities do not vary too much over time when there is no latent change (constraint (b)), whereas we expect that the probability of misplacing change, especially in ambiguous situations, increases with the length of the recall

Table 4
FLFS Observed Quarterly Transition Rates ($\times 100$), December 1990 to March 1992
(OM = Observed Mobility)

		EE	EU	EN	UE	UU	UN	NE	NU	NN	OM
D90-M91	WW	94.77	4.25	0.98	24.53	72.40	3.07	0.98	0.66	98.36	5.08
	BW	91.50	4.86	3.64	31.60	56.84	11.56	4.40	2.10	93.50	10.16
M91-J91	WW	96.03	3.02	0.95	23.21	74.32	2.47	1.28	0.68	98.04	4.54
	BW	91.48	4.63	3.89	35.01	54.20	10.79	4.84	2.14	93.02	12.04
J91-S91	WW	94.29	3.94	1.77	20.93	78.29	0.78	4.71	2.95	92.34	7.85
S91-D91	WW	93.73	4.48	1.79	23.63	74.89	1.48	3.22	1.65	95.13	7.23
D91-M92	WW	93.90	4.80	1.30	21.67	76.74	1.59	1.70	0.59	97.71	5.91

period. Constraints under (d) aim at catching the telescoping effect.

Figure 5 gives the path diagram of the estimated model.

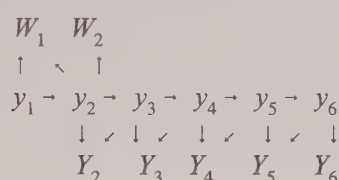


Figure 5. Path Diagram of a Modified Lisrel Model for Six Measurement Points and Two Indicators for One Latent Variable

Table B.1 in Appendix B reports the pattern of restrictions on response probabilities, (a) to (e); it shows which parameters are set equal, and which are fixed to 0, in order to introduce into the basic model, as defined by equations (B1) to (B6), the above constraints.

The final model has been selected after comparing a sequence of models, as can be seen from Table 5.

Table 5
Model Selection (EM = Estimated Mobility)

MODEL	L^2	df	ΔL^2	p-value cond. test	EM
A	2509.5759	2124			5.424
A1	3450.1716	2154	940.5957	0	4.918
A2	3849.9470	2178	399.7754	0	5.798
B	816.1620	2076			5.888
B1	855.2282	2094	39.0662	0.01	5.818
B2	864.9657	2106	9.7375	0.40	5.906
B3	879.5996	2121	14.6339	0.10	6.252

We started the analysis by estimating a model based on the ICE assumption (model A in the table), which, as expected, shows a bad fit.

The following models (A1 and A2) are based on the work by Magnac and Visser (1995). These authors consider monthly transitions over a period longer than ours (from January 1989 to March 1992), but on the same sample of individuals. They assume that the labour force state in the interview month is correctly reported, while the probability of making mistakes increases with the distance between the

reference month and the time of interview, according to a deterministic function of time. Response probabilities are assumed to be constant over the survey waves, and true transitions are assumed to follow a first order stationary Markov chain. Our model A1 is a less restricted version of Visser and Magnac's model – no stationarity assumption is made, applied to quarterly transitions from December 1990 to March 1992. Our model A2 adds to model A1, the hypothesis of first order stationarity at the latent level. Both models perform quite badly, and (from column EM), we see that, on average, they correct the observed labour market towards stability: a result which contradicts the evidence on the effects of classification errors in retrospective surveys.

Model B introduces correlation among classification errors, by letting each indicator to depend on the true transition that occurred between times t and $t+1$; moreover, it encompasses constraint (a). The fit increases dramatically (see L^2). All subsequent models are nested in model B, and additional restrictions may be evaluated by a conditional test. Model B1 introduces constraints under (b); model B2 the additional constraints under (c); and model B3 is our final model.

Table 6 presents estimated transition rates with our best fitting model. The French labour market is corrected towards a greater mobility. The average estimated mobility amounts to 6.252%. Moreover, estimated response probabilities show a pattern consistent with the notion, that the probability of making mistakes gets bigger, the longer the recall period.

ACKNOWLEDGEMENTS

Research for this paper was supported by grants from CNR n. 94.02242.CT10 and MURST n. 02.09.02.110 and n. 02.09.02.124. We are grateful to Michael Visser and Thierry Magnac for providing us with anonymized individual data from the French Labour Force Survey. A preliminary version of this paper was presented at the IASS/IAOS Satellite Meeting on Longitudinal Studies, Jerusalem, August 27-31, 1997, where we benefited from comments and discussion. Detailed criticisms and suggestions from two referees are especially acknowledged.

Table 6
FLFS Estimated Quarterly Transition Rates ($\times 100$), December 1990 to March 1992
(EM = Estimated Mobility)

	EE	EU	EN	UE	UU	UN	NE	NU	NN	EM
D90-M91	94.85	4.48	0.67	12.70	66.28	21.02	1.09	1.55	97.36	6.27
M91-J91	95.65	1.37	2.98	28.43	62.35	9.22	3.61	1.48	94.91	7.49
J91-S91	93.71	4.25	2.04	14.88	82.50	2.62	4.11	3.49	92.40	7.70
S91-D91	98.32	1.67	0.01	15.42	83.75	0.83	3.80	0.47	95.73	4.24
D91-M92	93.23	5.02	1.75	9.99	88.65	1.36	2.07	1.28	96.65	5.56

APPENDIX A

Final Model Specification for the SIPP Data, in Terms of Conditional Probabilities

(1) Basic model decomposition

$$z_g = P(G = g) \quad (A1)$$

$$\pi_1^{j_1} = P(y_1 = j_1) \quad (A2)$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t | y_{t-1} = j_{t-1}) \quad t = 2, 3, 4 \quad (A3)$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1} g} = P(Y_t = l_t | y_t = j_t, Y_{t+1} = l_{t+1}, y_{t+1} = j_{t+1}, G = g) \quad t = 1, 2, 3 \quad (A4)$$

$$q_{y4}^{l_4 j_4 g} = P(Y_4 = l_4 | y_4 = j_4, G = g) \quad (A5)$$

$$q_{wt}^{m_t j_t m_{t+1} j_{t+1} g} = P(W_t = m_t | y_t = j_t, W_{t+1} = m_{t+1}, y_{t+1} = j_{t+1}, G = g) \quad t = 1, 2, 3 \quad (A6)$$

$$q_{w4}^{m_4 j_4 g} = P(W_4 = m_4 | y_4 = j_4, G = g) \quad (A7)$$

g varies over 1, 2, 3 and 4; l_t and j_t , $t = 1, 2, 3, 4$, vary over the categories E, U and N, m_t , $t = 1, 2, 3, 4$, vary over the categories 'Job' and 'No Job'.

(2) Constraints on conditional probabilities

$$q_{yt}^{l_{t+1} j_{t+1} l_t j_t g} = P(Y_t = l_{t+1} | y_t = j_t, Y_{t+1} = l_{t+1}, y_{t+1} = j_{t+1}, G = g) = 1 \quad \text{for } l_{t+1} \neq j_{t+1} \quad t = 1, 2, 3 \quad (A8)$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1} g} = P(Y_t = l_t | y_t = j_t, G = g) \quad \text{for } l_{t+1} = j_{t+1} \text{ and } t = 1, 2, 3 \quad (A9)$$

$$q_{wt}^{m_{t+1} j_{t+1} m_t j_t g} = P(W_t = m_{t+1} | y_t = j_t, W_{t+1} = m_{t+1}, y_{t+1} = j_{t+1}, G = g) = 1 \quad \text{for } m_{t+1} \neq j_{t+1} \text{ and } t = 1, 2, 3 \quad (A10)$$

$$q_{wt}^{m_t j_t m_{t+1} j_{t+1} g} = P(W_t = m_t | y_t = j_t, G = g) \quad \text{for } m_{t+1} = j_{t+1} \text{ and } t = 1, 2, 3 \quad (A11)$$

$$q_{t=4}^{g=1} = q_{t=3}^{g=4} = q_{t=2}^{g=3} = q_{t=1}^{g=2} \quad (A12)$$

$$q_{t=4}^{g=2} = q_{t=3}^{g=1} = q_{t=2}^{g=4} = q_{t=1}^{g=3} \quad (A13)$$

$$q_{t=4}^{g=3} = q_{t=3}^{g=2} = q_{t=2}^{g=1} = q_{t=1}^{g=4} \quad (A14)$$

$$q_{t=4}^{g=4} = q_{t=3}^{g=3} = q_{t=2}^{g=2} = q_{t=1}^{g=1} \quad (A15)$$

APPENDIX B

Final Model Specification for the FLFS Data, in Terms of Basic Model Decomposition and Pattern of Restrictions on Parameters

(1) Basic model decomposition

$$\pi_1^{j_1} = P(y_1 = j_1) \quad (B1)$$

$$\pi_t^{j_t j_{t-1}} = P(y_t = j_t | y_{t-1} = j_{t-1}) \quad t = 2, 3, 4, 5 \quad (B2)$$

$$q_{yt}^{l_t j_t l_{t+1} j_{t+1}} = P(Y_t = l_t | y_t = j_t, y_{t+1} = j_{t+1}) \quad t = 2, 3, 4, 5 \quad (B3)$$

$$q_{y6}^{l_6 j_6} = P(Y_6 = l_6 | y_6 = j_6) \quad (B4)$$

$$q_{wt}^{m_1 j_1 j_2} = P(W_1 = m_1 | y_1 = j_1, y_2 = j_2) \quad (B5)$$

$$q_{w2}^{m_2 j_2} = P(W_2 = m_2 | y_2 = j_2) \quad (B6)$$

j_t , l_t and m_t vary over E, U and N.

(2) Pattern of restrictions on response probabilities

Table B.1
Month of Observation

Probability of observing a state given a latent transition	March 91	June 91 & Sept. 91	Dec. 90 & Dec. 91
E _{lee}	1	1	1
U _{lee}	2	2	2
N _{lee}	3	3	3
E _{leu}	F	F	F
U _{leu}	F	F	F
N _{leu}	*	*	*
E _{len}	F	F	F
U _{len}	*	*	*
N _{len}	F	F	F
E _{lue}	F	F	F
U _{lue}	F	F	F
N _{lue}	*	*	*
E _{luu}	4	4	4
U _{luu}	5	5	5
N _{luu}	6	6	6
E _{lun}	*	*	*
U _{lun}	F	F	F
N _{lun}	F	F	F
E _{lne}	F	F	F
U _{lne}	F	F	F
N _{lne}	F	F	F
E _{lnu}	*	*	*
U _{lnu}	F	F	F
N _{lnu}	F	F	F
E _{lnn}	7	7	7
U _{lnn}	8	8	8
N _{lnn}	9	9	9

Equal numbers indicate response probabilities fixed to be equal.

* indicates a probability fixed to 0.

F indicates a free parameter.

REFERENCES

- ABOWD, J.M., and ZELLNER A. (1985). Estimating gross labour force flows. *Journal of Business and Economics Statistics*, 3, 254-283.
- BASSI, F., CROON, M., HAGENAARS, J., and VERMUNT, J. (1995). Estimating Latent Turnover Tables When Data are Affected by Correlated and Uncorrelated Classification Errors. WORC PAPER 95.12.25/7, Tilburg University.
- BERNARD, H.R., KILLWORTH, P., KRONENFELD, D., and SAILER, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495-517.
- CHAKRABARTY, R.P., and WILLIAMS, T.R. (1989). Time-in-sample biases in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 309-314.
- CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.
- CITRO, C.F., and KALTON, G. (1993). *The Future of the SIPP*. Washington, D.C.: National Academy Press.
- DUNCAN, G., and KALTON, G. (1987). Issues of design and analysis of surveys across time. *International Statistical Review*, 55, 97-117.
- GOODMAN, L.A. (1973). The analysis of a multidimensional contingency table when some variables are posterior to the others. *Biometrika*, 60, 179-192.
- GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 79, 1178-1259.
- GOODMAN, L.A. (1981). Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. In *Sociological Methodology*, (Ed. S.Leinhardt), 193-293. San Francisco: Jossey Bass.
- HABERMAN, S.J. (1978). *Analysis of Qualitative Data. Vol.1. Introductory Topics*. New York: Academic Press.
- HAGENAARS, J.A. (1988). Latent structure models with direct effects between the indicators, local dependence models. *Sociological Methods and Research*, 16, 379-405.
- HAGENAARS, J.A. (1990). *Categorical Longitudinal Data: Log-linear, Panel, Trend and Cohort Analysis*. Newbury Park: Sage.
- HAGENAARS, J.A. (1997). Categorical Causal Modeling: Directed Loglinear Models With Latent Variables. WORC PAPER 97.04.002/7, Tilburg University.
- HUBBLE, D.L., and JUDKINS, D.R. (1989). Measuring the Bias in Gross Flows in the Presence of Autocorrelated Response Errors. SIPP Working Paper No. 8712, U.S. Bureau of the Census.
- HUSSMAN, R., MEHRAN, F., and VERMA, M. (1990). *Surveys of Economically Active Population, Employment and Underemployment: an ILO Manual on Concepts and Definitions*. Geneva: ILO.
- JORESÖG, K.G., and SÖRBOM, D. (1988). *Lisrel 7: A Guide to the Program and Applications*. Chicago: SPSS INC.
- KALTON, G., and CITRO, C.F. (1993). Panel surveys: adding the fourth dimension. *Survey Methodology*, 19, 205-215.
- KALTON, G., and MILLER, M.W. (1991). The seam effect with social security income in the SIPP. *Journal of Official Statistics*, 7, 235-245.
- LANGHEINE, R., PANNEKOEK, J., and van de POL, F. (1995). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24, 492-516.
- LAZARSFELD, P.F., and HENRY, N.W. (1968). *Latent Structure Analysis*. New York: Houghton Mufflin.
- MAGNAC, T., and VISSER, M. (1995). Transition Models With Measurement Errors. Working Paper, Institut National de la Recherche Agronomique (INRA), Paris.
- MARQUIS, K.H., and MOORE, J.C. (1989). Some response errors in SIPP with thoughts about their effects and remedies. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 381-386.

- MARTINI, A. (1988). Retrospective versus panel data in estimating labour force gross flows: comparing SIPP and CPS. *Proceedings of the Social Science Section, American Statistical Association*, 109-114.
- MARTINI, A. (1989). Seam effect, recall bias, and the estimation of labour force transition rates from SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 387-392.
- MENG, X.L., and RUBIN, D.B. (1993). Maximum likelihood estimation via ECM algorithm: a general framework. *Biometrika*, 80, 267-278.
- MCCORMICK, M., BUTLER, D., and SINGH, R. (1992). Investigating time-in-sample effects for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 554-559.
- O'MUIRCHEARTAIGH, C. (1996). Measurement errors in panel surveys: implications for survey design and for survey instruments. *Proceedings of the Scientific Reunion of the Italian Statistical Society*, 1, 207-218. Rimini: Maggioli.
- PFEFFERMAN, D., SKINNER, C.J., and HUMPHREYS, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, Series A*, 161, 13-32.
- POTERBA, J.M., and SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.
- SINGH, A.C., and RAO, J.N.K. (1995). On the adjustment of gross flows estimates for classification errors with application to data from the Canadian Labor Force Survey. *Journal of the American Statistical Association*, 90, 1-11.
- SKINNER, C.J., and TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 3, 391-405.
- U.S. DEPARTMENT OF COMMERCE (1991). *SIPP User's Guide*. Washington D.C.
- van de POL, F., and LANGEHEINE, R. (1990). Mixed Markov latent class models. In *Sociological Methodology*, (Ed. C.Clogg), 213-247. Oxford: Blackwell.
- van de POL, F., and LANGEHEINE, R. (1992). Analysing Measurement Error in Quasi-experimental Data: An Application of Latent Class Models to Labour Market Data. Working Paper of the European Scientific Network on household Panel Studies, 57, Colchester, University of Essex.
- van de POL, F., and LANGEHEINE, R. (1997). Separating change and measurement error in panel surveys with an application to labour market data. In *Survey Measurement and Process Quality*, (Ed. L.Lyberg *et al.*), 671-688. New York: Wiley.
- VERMUNT, J.K. (1993). Log-linear and event history analysis with missing data using the EM algorithm. WORC PAPER 93.09.015/7, Tilburg University.
- YOUNG, N. (1989). Wave seam effects in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 393-398.

Estimating Labour Force Gross Flows From Surveys Subject to Household-level Nonignorable Nonresponse

PAUL S. CLARKE and RAY L. CHAMBERS¹

ABSTRACT

Measurement of gross flows in labour force status is an important objective of the continuing labour force surveys carried out by many national statistics agencies. However, it is well known that estimation of these flows can be complicated by nonresponse, measurement errors, sample rotation and complex design effects. Motivated by nonresponse patterns in household-based surveys, this paper focuses on estimation of labour force gross flows, while simultaneously adjusting for nonignorable nonresponse. Previous model-based approaches to gross flows estimation have assumed nonresponse to be an individual-level process. We propose a class of models that allow for nonignorable household-level nonresponse. A simulation study is used to show, that individual-level labour force gross flows estimates from household-based survey data, may be biased and that estimates using household-level models can offer a reduction in this bias.

KEY WORDS: Gross flows; Household-based surveys; Nonignorable nonresponse.

1. INTRODUCTION

Labour force gross flows are typically defined as transitions over time between the three major labour force states, employed, unemployed and economically inactive. Gross flows estimates are an important tool in the study of labour force dynamics (for example, see Vanski 1985). Large-scale on-going surveys such as the British Labour Force Survey and the U.S. Current Population Survey, provide data for gross flows estimation. However, non-response, measurement error, sample rotation and complex design effects, affect gross flows estimation from these surveys. A discussion of these and other factors affecting gross flows estimation, is given in Hogue (1985). Here we focus on the problem of nonresponse.

We assume that a nonresponse mechanism leads to the observed data being incomplete. If the probability of not responding depends on the missing data, then the non-response mechanism is nonignorable (Rubin 1976). The model-based approach to analysing incomplete survey data, is detailed in Little (1982). Model-based approaches to the estimation of labour force gross flows, involve modelling both the labour force flows and the nonresponse mechanism, and simultaneously fitting both models to the incomplete data. Examples of such models are given in Stasny and Fienberg (1985), Stasny (1986) and, for nonignorable nonresponse, in Little (1985). We call these individual-level models, because individuals are modelled as responding or not responding, independently of other sampled individuals.

Both the Labour Force Survey and the Current Population Survey, are examples of household-based surveys, that is, surveys based on a random sample of households, rather than individuals. Household-based surveys can lead to correlated nonresponse behaviour

within households. For example, in the Current Population Survey, a single household member (usually the head-of-household) acts as a proxy for the other household members; thus, if the chosen household member is a non-respondent, so are other household members. It follows that, due to correlated within-household nonresponse behaviour, individual-level nonresponse models are unsuitable for the estimation of labour force gross flows, using household-based survey data.

In this paper, we propose a class of models for individual-level labour force flows, and household-level nonresponse, that account for correlated within-household nonresponse behaviour. A number of plausible nonresponse models that are estimable from the observed data, both ignorable and nonignorable, are also presented. We then simulate household-based survey data, using these household-level models, to demonstrate the potential utility of our approach: first, individual-level labour force gross flows estimates are shown to be biased, when fitted to household-based survey data; and second, the bias of individual-level and household-level gross flows estimates are compared, to show the advantages of fitting household-level models to household-based survey data. To conclude, we summarise the findings of our simulation studies and discuss ideas for further research in this area.

2. A MODEL FOR HOUSEHOLD-LEVEL NONRESPONSE

2.1 The Data

A gross flow is the probability or frequency of individuals in the population, making a state transition between two points in time, t_1 and t_2 ($t_1 < t_2$). Labour force gross flows refer to transitions between the three main

¹ Paul S. Clarke and Ray L. Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

labour force states: 1 = 'employed', 2 = 'unemployed' and 3 = 'not in labour force', where the last category refers to economically inactive individuals, such as retired individuals and students. Let S denote a simple random sample of households, indexed by h . Within household h , there are n_h eligible individuals, of which $n_h(ab)$ have labour force flow (a, b) between t_1 and t_2 , where $\sum_{a,b} n_h(ab) = n_h$, and $a, b = 1, 2, 3$. We refer to $\{n_h(ab)\}$ as the complete data, that is, the frequencies that would be observed in the absence of nonresponse.

Table 1 shows the complete labour force flows data for household h as a 3×3 contingency table. If h responds at both times, the observed data are the cells of this 2-way table. However, if the household does not respond at t_1 or t_2 , the observed data correspond to the margins of the table: $n_h(1+)$, $n_h(2+)$, $n_h(3+)$ are the observed data if h responds at t_1 , but does not respond at t_2 ; and $n_h(+1)$, $n_h(+2)$, $n_h(+3)$ are the observed data if h responds at t_2 but does not respond at t_1 . (An index replaced by '+' denotes summation over all levels of that index.) Furthermore, if h does not respond at both t_1 and t_2 , the observed data is the household size, n_h , which we take to be known and fixed between t_1 and t_2 .

Table 1

Complete Labour Force Flows Data for Household h

Status		t_2			
		1	2	3	
t_1	1	$n_h(11)$	$n_h(12)$	$n_h(13)$	$n_h(1+)$
	2	$n_h(21)$	$n_h(22)$	$n_h(23)$	$n_h(2+)$
	3	$n_h(31)$	$n_h(32)$	$n_h(33)$	$n_h(3+)$
		$n_h(+1)$	$n_h(+2)$	$n_h(+3)$	n_h

2.2 Model Specification

It is inappropriate to treat the nonresponse behaviour of individuals within a household as independent, in household-based surveys. In the Labour Force Survey, for example, one eligible household member determines whether the household can be interviewed. Therefore, if no eligible individual can be contacted, each household individual is a nonrespondent. To construct a model for household-level nonresponse, we take the ideas behind individual-level nonresponse and extend them to the household, by considering a household to be an entity with its own nonresponse flow between t_1 and t_2 . To allow for nonignorable nonresponse, the probability of a household nonresponse flow is modelled as a function of its individual labour force flows, as shall now be described.

Let $N_h = (N_h(11), N_h(21), \dots, N_h(33))$ be the random vector of labour force flows frequencies for household h , where $N_h(ab)$ is the random variable, whose outcome corresponds to the number of individuals with labour force flow (a, b) , $a, b = 1, 2, 3$. Further, denote the random vector

for the nonresponse flow of household h by $R_h = (R_{h1}, R_{h2})$, where

$$R_{hj} = \begin{cases} 1, & \text{if household responds at } t_j \\ 0, & \text{otherwise} \end{cases}$$

is the nonresponse status random variable for h at t_j , $j = 1, 2$. The realisations of these random quantities are denoted by \mathbf{n}_h and \mathbf{r}_h . We now assume that \mathbf{n}_h and \mathbf{r}_h are known, and write the joint probability of N_h and R_h as

$$\Pr(N_h = \mathbf{n}_h, R_h = \mathbf{r}_h) = \Pr(N_h = \mathbf{n}_h) \Pr(R_h = \mathbf{r}_h | N_h = \mathbf{n}_h),$$

where $\Pr(N_h = \mathbf{n}_h)$ is the labour force flows model, and $\Pr(R_h = \mathbf{r}_h | N_h = \mathbf{n}_h)$ is called the nonresponse flows model.

The labour force flows model is taken to be multinomial, with probability function

$$\Pr(N_h = \mathbf{n}_h; \boldsymbol{\omega}) = n_h! \prod_{a,b} \frac{\omega(ab)^{n_h(ab)}}{n_h(ab)!}, \quad (1)$$

where $\omega(ab) > 0$ is the probability of an individual having labour force flow (a, b) and $\sum_{a,b} \omega(ab) = 1$. The vector of labour force flows parameters is denoted by $\boldsymbol{\omega} = (\omega(11), \omega(21), \dots, \omega(33))$, of which 8 are free. The assumption of multinomial sampling in (1), implies that individuals' labour force flows behaviour, is independent within households, and that households are homogeneous with respect to their labour force flows behaviour. These assumptions are unrealistic, but (1) can easily be extended to a more realistic model for the labour force flows, as we discuss in Section 4.

The probability of household h having nonresponse flow (u, v) , is taken to be

$$\begin{aligned} \pi(uv | \mathbf{n}_h) &= \Pr(R_h = (u, v) | N_h = \mathbf{n}_h; \boldsymbol{\psi}) \\ &= \frac{1}{n_h} \sum_{a,b} n_h(ab) \psi(uv | ab), \end{aligned} \quad (2)$$

for $u, v = 0, 1$, namely, a weighted average of the non-response model parameters. By setting $n_h = 1$, it can be seen that $\psi(uv | ab) > 0$ is the probability of a household of size one (*i.e.*, an individual) having nonresponse flow (u, v) , given it has labour force flow (a, b) . Thus, $\sum_{u,v} \psi(uv | ab) = 1$ and $\boldsymbol{\psi} = (\psi(11 | 11), \psi(01 | 11), \dots, \psi(00 | 33))$ is the vector of nonresponse parameters, of which 27 are free.

Before defining the likelihood function for the complete data, partition S into 4 mutually exclusive and exhaustive subsets

$$S = S_{11} \cup S_{01} \cup S_{10} \cup S_{00},$$

where $S_{uv} = \{h: r_h = (u, v)\}$ is the subset of households with nonresponse flow (u, v) . Thus, since S is a simple random sample of households, the likelihood function for the complete data is

$$L(\omega, \psi; \{n_h, r_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; n_h, (u, v)), \quad (3)$$

where $L_h(\omega, \psi; n_h, (u, v))$ is the contribution of household $h \in S_{uv}$ to the likelihood, the product of (1) and (2).

2.3 Model Fitting

2.3.1 Maximum Likelihood Estimation

Since the complete data are unavailable, (3) must be modified to give the likelihood based on the observed data. Denote the observed data by $\{n_h^*\}$. As discussed in Section 2.1, the observed data for households that respond at t_1 and t_2 , is the full cross-classification in Table 1, namely, $n_h^* = n_h$. Similarly, if $h \in S_{10}$ then $n_h^* = (n_h(1+), n_h(2+), n_h(3+))$; if $h \in S_{01}$ then $n_h^* = (n_h(+1), n_h(+2), n_h(+3))$; and if $h \in S_{00}$, then $n_h^* = n_h$.

The contribution of household $h \in S_{uv}$ to the observed data likelihood, is obtained by summing $L_h(\omega, \psi; n_h, (u, v))$ over all possible values that the full 3×3 cross-classification of labour force flows can take, given the observed margin. Representing this set of tables by $n_h: n_h^*$, the observed data likelihood for S is

$$L(\omega, \psi; \{n_h^*, r_h\}) = \prod_{u,v} \prod_{h \in S_{uv}} \sum_{n_h: n_h^*} L_h(\omega, \psi; n_h, (u, v)). \quad (4)$$

Model fitting requires calculating (4) at each stage of an iterative optimization process. This is computationally intensive, because the complete data likelihood function must be summed explicitly over the missing data. For example, the observed data for $h \in S_{10}$ is $n_h^* = (n_h(1+), n_h(2+), n_h(3+))$ and the likelihood contribution of this household to the observed data likelihood is

$$\sum_{n_h: n_h^*} L_h(\omega, \psi; n_h, (1, 0)).$$

To explicitly calculate this contribution, each 3×3 complete data table n_h for fixed n_h^* is generated and $L_h(\omega, \psi; n_h, (1, 0))$ evaluated for each. For household size $n_h = 5$, there are at least 21 and at most 108 possible tables, depending on the values in the fixed margin; for $n_h = 15$, a very large household size, the respective numbers are 136 and 9,261. A similar procedure is used for $h \in S_{01}$, except here $n_h^* = (n_h(+1), n_h(+2), n_h(+3))$ is the fixed margin. If $h \in S_{00}$, then no data about labour force status are observed, only the household size n_h . So each 3×3 table with total n_h must be generated, and the likelihood function calculated for each: for $n_h = 5$ there are 1,287 tables and for $n_h = 15$ there are 490,314. It is not infeasible, in terms of computer run-time, to calculate such sums directly. The number of

explicit calculations can be reduced, by recognising that each household is defined only by its observed labour force flows frequencies and nonresponse flow. Thus, summation over the missing data need only be performed once for a household with a particular nonresponse flow and labour force flows frequencies; the contribution of this household to the likelihood is then raised to the power of the number of similarly defined households in S .

2.3.2 Parameter Estimability

If we fix $n_h = 1$ for all h , the complete data have no household structure, and form a 4-way table cross-classified by labour force status and nonresponse status at t_1 and t_2 . The observed data log-likelihood (4) is now equivalent to that of the individual-level models in Stasny and Fienberg (1985), Little (1985) and Stasny (1986). For these models, estimability requires that the number of model parameters does not exceed 15 (one for each observed table cell, less one for the multinomial sampling constraint). Hence, (ω, ψ) are inestimable because there are $8 + 27 = 35$ free parameters. Since interest is focused on the labour force gross flows probabilities, ω , it is necessary to constrain ψ to ensure estimability.

When $n_h > 1$, determining parameter estimability is more difficult, because (4) has a complicated closed-form expression. Fitzmaurice, Laird and Zahner (1996) use a numerical method to determine estimability, that involves showing that the information matrix is non-singular in the neighbourhood of the maximum likelihood estimate. However, not only is this impractical for problems of a high dimension, but evaluating the information matrix for the household-level model, is particularly difficult in this case. Instead, we adopt a pragmatic approach for determining parameter estimability: first, we restrict attention to models that satisfy the necessary condition for estimability when $n_h = 1$; and second, different starting values are used to for each fit. If the different starting values reveal a non-unique maximum likelihood estimate, or any parameter estimate is unchanged from its starting value then the model parameters are taken to be inestimable.

2.4 Nonresponse Models

To enable parameter estimates to be obtained from the observed data, θ and ψ must be constrained in accordance with assumptions about the nonresponse mechanism. The nonresponse parameters are interpreted as individual nonresponse probabilities, but within the household framework established thus far, it is inappropriate to talk about individuals not responding. However, in reality, it is individuals within households that determine a household's nonresponse flow, not the household itself. Therefore, constraints are placed on the nonresponse parameters at the individual level, that apply at the household level through the functional dependence of $\pi(uv | n_h)$ on ψ in (2). For example, if the nonresponse parameters are constrained such that $\psi(uv | ab) = \psi(uv)$ for all a, b , then the household nonresponse mechanism is ignorable, because household

nonresponse flows are independent of the labour force flows.

We now present four models for the nonresponse mechanism, two of which are ignorable, and two nonignorable.

– Ignorable models.

– Model I_A : Constant nonresponse probability,

$$\psi(uv|ab) = \lambda^{1-u}(1-\lambda)^u \times \lambda^{1-v}(1-\lambda)^v,$$

which has 1 parameter, λ , the probability of an individual not responding;

– Model I_B : Independent of labour force status, but different nonresponse probabilities, at t_1 and t_2 ,

$$\psi(uv|ab) = \lambda^{1-u}(1-\lambda)^u \times \theta^{1-v}(1-\theta)^v,$$

which has 2 parameters, λ, θ , the probabilities of nonresponse at t_1 and t_2 , respectively.

– Nonignorable models.

– Model N_A : The nonresponse distributions at t_1 and t_2 are independent but depend on labour force status at t_1 and t_2 , respectively,

$$\psi(uv|ab) = \lambda(a)^{1-u}(1-\lambda(a))^u \times \theta(b)^{1-v}(1-\theta(b))^v$$

which has 6 parameters, $\lambda = (\lambda(1), \lambda(2), \lambda(3))$ and $\theta = (\theta(1), \theta(2), \theta(3))$, where $\lambda(a)$ is the probability of not responding at t_1 , given labour force status a at t_1 , and $\theta(b)$ that at t_2 , given labour force status b at t_2 ;

– Model N_B : The nonresponse distributions at t_1 and t_2 depend on labour force status at t_1 and t_2 respectively, *i.e.*, a first-order Markov process. Unlike N_A , the nonresponse distributions at t_1 and t_2 are dependent: if the nonresponse status at t_1 is 1, then the nonresponse distribution at t_2 is the same as at t_1 ; but if the nonresponse status at t_1 is 0, the nonresponse distributions are distinct,

$$\psi(uv|ab) = \lambda(a)^{1-u}(1-\lambda(a))^u$$

$$\times \begin{cases} \lambda(b)^{1-v}(1-\lambda(b))^v, & \text{if } u = 1, \\ \theta(b)^{1-v}(1-\theta(b))^v, & \text{if } u = 0, \end{cases}$$

for $a, b = 1, 2, 3$ and $u, v = 0, 1$. Under model I_A , there are a total of $8 + 1 = 9$ free parameters, satisfying the necessary condition for estimability of an individual-level model. Models I_B , N_A and N_B have 10, 14 and 14 free parameters, respectively, and so also satisfy the necessary condition for estimability.

3. SIMULATION STUDY

3.1 Simulation Procedure

We used a simulation study to investigate the consequences of failing to account for the household structure of

household-based survey data, and to compare labour force gross flows estimates for individual-level and household-level models. For this purpose, household-based survey data was generated using Monte Carlo sampling. Each sample data set consisted of 10,000 individuals arranged into households of size $n_h = k$ for all h . Within each household, labour force flows were generated from (1), and the nonresponse flow was generated from (2), under one of models N_A or N_B . The data were made incomplete by collapsing each complete labour force flows data table, to be consistent with the household nonresponse flow. In total, 1,000 independent data sets were generated in this way.

The population parameters used to generate the labour force flows are shown in the following table:

		b			
		1	2	3	
a	$\omega(ab)$	1	0.43	0.245	0.035
	2	0.02	0.160	0.01	
	3	0.015	0.035	0.05	

This is clearly a population in recession, since the probability of moving from being employed to unemployed is very large ($\omega(12) = 0.245$). Under models N_A and N_B , the population parameters are

		i		
		1	2	3
$\lambda(i)$		0.2	0.8	0.5
$\theta(i)$		0.5	0.2	0.8

It should be noted that these parameter values do not represent realistic nonresponse flows behaviour, they were chosen for the purpose of illustrating this methodology. However, this does not affect the general conclusions of the paper, which are also relevant for realistic values of the true nonresponse probabilities.

3.2 Simulation Results

Estimates for individual-level models are obtained by fitting (4) with $n_h = 1$ to each incomplete data set. Figure 1 summarises the sampling distributions of the individual-level maximum likelihood estimate of $\omega(12)$, $\hat{\omega}(12)$, for nonresponse models I_A , I_B , N_A and N_B (estimates for ignorable models I_A and I_B are included together, because both yield the same estimates of the labour force flows). The vertical lines represent the intervals between the 2.5-percentile and the 97.5-percentile of each estimate's sampling distribution, and the bold point represents its median. There are three distributions obtained for each individual-level estimate: the left-most distribution is that when the household size is $k = 1$, *i.e.*, the simulated data

have no household structure; and reading from left to right, the next two distributions are those obtained when the household size is $k = 2$ and $k = 5$, respectively. The solid horizontal line denotes the true flow probability, $\omega(12) = 0.0245$. The behaviour of the sampling distribution of $\hat{\omega}(12)$ in this study, reflects that of the other labour force gross flows estimates.

Figure 1a summarises the sampling distributions when N_A is the true model. If the fitted individual-level model is I_A , I_B or N_B , the labour force gross flows estimates have large biases, whatever the household size. As would be expected, the median estimate for correct model N_A , is unbiased if $k = 1$ and a small bias is apparent for $k = 2$ and $k = 5$ (although this bias is smaller for $k = 5$ than $k = 2$). Bias reduction with increasing k is also apparent for individual-level estimates I_A , I_B and N_B . This behaviour is unexpected, since it seems natural to expect the bias of the individual-level estimates, to increase with the household size. The results are slightly different in Figure 1b when N_B is true. Here the estimate for individual-level model N_B , becomes more biased as k increases, but the bias decreases for mis-specified individual-level models I_A , I_B and N_A . Furthermore, the misspecified estimates for I_A and I_B have a small bias, when compared to those for misspecified model N_A . These results are discussed in Section 3.3.

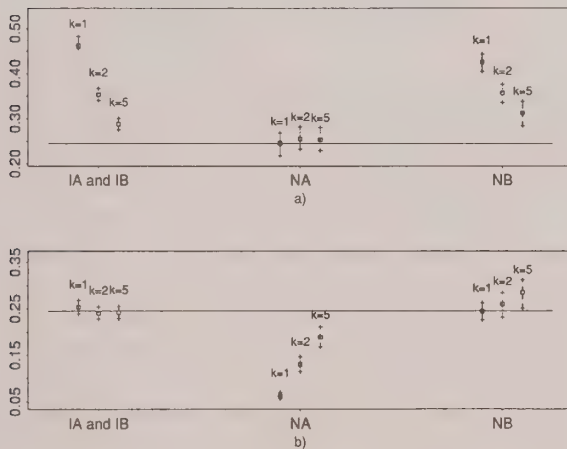


Figure 1. Sampling Distribution of $\hat{\omega}(12)$ for Individual-Level Models I_A , I_B , N_A and N_B When the True Nonresponse Model is a) N_A and b) N_B and the Household Size is $k = 1, 2, 5$.

A comparison of the median estimates of $\omega(12)$ for the fitted individual-level and household-level models when N_B is true, is presented in Figure 2. There are four sampling distributions associated with each model: the first two represent those from fitting an individual-level nonresponse model, and a household-level nonresponse model, when the household size is $k = 2$; and similarly, the next two distributions are those when the household size is 5.

For a particular pair of individual-level and household-level sampling distributions, it can be seen that the household-level estimate is less biased than its equivalent individual-level estimate, and the spread of each household-level sampling distribution, is narrower. The exception to this, is when fitting model I_A , where the household-level and individual-level distributions are identical. This equality occurs because the observed data likelihood for the individual-level and household-level models, are equivalent when the nonresponse model is ignorable. Another feature is that, if the nonresponse model is correctly specified, the household-level estimates are unbiased.

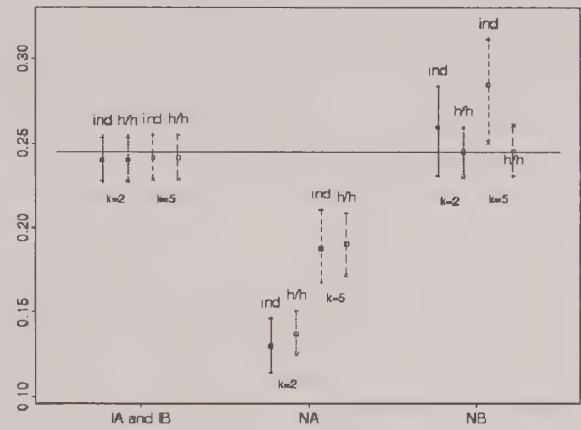


Figure 2. Sampling Distributions of $\hat{\omega}(12)$ for Individual-Level and Household-Level Models I_A , I_B , N_A and N_B When the True Nonresponse Model is N_B and the Household Size is $k = 2, 5$.

3.3 Summary

The estimates of the labour force gross flows under individual-level models, are never less biased than those of household-level models, when fitted to household-based survey data in our study. It should be noted, that if the true model is ignorable, it is unnecessary to utilise a household-level nonresponse model, because the individual-level and household-level models are equivalent. For example, if I_A is true, (2) reduces to $\lambda^{u+v}(1-\lambda)^{1-u-v}$, and (4) factorizes into two components, dependent on ω only and λ only; the factor dependent on ω can be shown to be equivalent to that for the individual-level model, and thus the labour force flows estimates are the same.

It appears, as the household size increases, that the bias of the labour force flows estimates decreases, if the true model is nonignorable. In fact, this result arises because we use (1) to generate the labour force flows, and not because the model estimates are unbiased for large n_h . To see why, consider the household formation process, used to generate

each Monte Carlo sample: as n_h increases, each household frequency tends to the same value, *i.e.*, $n_h(ab)$ converges to $n_h \omega(ab)$; hence,

$$\begin{aligned} \pi(uv | n_h) &\rightarrow \frac{1}{n_h} \sum_{a,b} n_h \omega(a, b) \psi(uv | ab) \\ &= \sum_{a,b} \omega(ab) \psi(uv | ab), \end{aligned}$$

which is independent of n_h , that is, the simulated household nonresponse mechanism is ignorable. Therefore, the labour force flows estimates are unbiased, because fitting the nonignorable models to the simulated data, yields parameter estimates that are consistent with ignorable nonresponse. To generate nonignorable household-level nonresponse, it is necessary to prevent $n_h(ab) \rightarrow n_h \omega(ab)$, by extending (1), to allow for differential labour force flows between households. Such extensions to the labour force flows model are discussed in Section 4.

Figure 1b) shows two anomalous results that contradict the above explanation, when N_B is the true model. First, the bias of individual-level model N_B 's estimate, increases as n_h increases. However, further simulations with household size $n_h = 10$, revealed that the individual-level estimate bias is zero. Thus, asymptotic ignorable nonresponse is also evident when N_B is true, but n_h must be large before its effect becomes apparent for individual-level model N_B . Second, the bias of the ignorable individual-level model estimates is small, almost zero, when N_B is true. This small bias reduces even further as n_h increases, in line with asymptotic ignorability, but we have yet to arrive at a satisfactory explanation as to why the ignorable models perform so well in this situation. Further study is necessary to investigate this finding.

4. DISCUSSION

In Sections 3 and 4, it is demonstrated by means of a simulation based study, that modelling household-level nonignorable nonresponse, when estimating labour force gross flows from household-based surveys, leads to reduced bias in the flows estimates, compared to those from individual-level models. If the nonresponse model is ignorable, it is unnecessary to use household-level models, because the individual-level and household-level models are equivalent. Furthermore, it is shown that controlling for household-level nonresponse does not necessarily remove all bias from the estimates of the labour force flows. Correct specification of the nonresponse model is still seen to be imperative, although taking the household structure of the data into account, may lead to a refinement of the flows estimates if the nonresponse model is misspecified. In particular, we show that household-level estimates are less biased than their equivalent individual-level estimates.

Our nonresponse model is an extension of the idea that nonresponse can depend upon the characteristics of a unit, in this case, the labour force flows of household members. Nonresponse in household-based surveys can occur for more than one reason, *e.g.*, refusal, non-contact, moving house or sample rotation. The current model can easily be extended to model more complex nonresponse patterns, by specifying the nonresponse indicator as a polytomous variable, and parameterizing the nonresponse model in accordance with the complex nonresponse patterns. It should also be noted, that we do not assume that the household-level model is an accurate representation of household nonresponse behaviour; rather, we assume that the household-level model, offers an approximation of within-household nonresponse dynamics.

An important problem, highlighted by the results from the simulation study, is our assumption that individual labour force flows behaviour is homogeneous within households. Clearly, this is an unrealistic assumption. The model is easily extended, by specifying the labour force flows and nonresponse flows probabilities, as regression models to accommodate individual-level, household-level, or higher level covariate information. For example, the labour force flows probabilities could be specified as a multinomial-logistic regression:

$$\log \left(\frac{\omega_{hi}(ab)}{\omega_{hi}(11)} \right) = \beta_0^{(ab)} + \beta_1^{(ab)} \mathbf{x}_{hi}^T,$$

where $\omega_{hi}(ab)$ denotes the probability of individual i in household h , making labour force flow (a, b) , \mathbf{x}_{hi} is a (row) vector of covariates, and $(\beta_0^{(ab)}, \beta_1^{(ab)})$ are the regression coefficients for multinomial-logit (a, b) . However, fitting these models requires conditional independence assumptions to be made, about the relationship between the distributions of the covariates, the labour force flows and the nonresponse flows, because the covariate information may be missing for nonresponding households. An alternative solution, is to allow for heterogeneous between household labour force flows, using random effects, by making assumptions about the distribution of between household differences. Fitting these models is also complicated and would require, for example, a Markov chain Monte Carlo procedure to perform the necessary integration. If S is not a simple random sample, auxiliary design variables can be incorporated into the fitting process, using the regression framework just described.

ACKNOWLEDGEMENTS

The work of Paul Clarke on this paper was funded by an Economic and Social Research Council studentship (award no. R00429614273); the work of Ray Chambers was funded by the contract between the Office for National

Statistics and the University of Southampton for the provision of research services in statistical methodology. Both authors would like to thank the referees, whose comments and practical advice helped make the final version of the manuscript considerably more readable.

REFERENCES

- FITZMAURICE, G.M., LAIRD, N.M., and ZAHNER, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-107.
- HOGUE, C.R. (1985). History of the problems encountered in estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 1-8.
- LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.
- LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin of the International Statistical Association*, 15, 1-15.
- RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 81, 42-47.
- STASNY, E.A., and FIENBERG, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 25-39.
- VANSKI, J.E. (1985). Uses of gross change data in assessing demographic labor market dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 9-12.

Longitudinal Analysis of Swiss Labour Force Survey Data by Multivariate Logistic Regression

PAUL-ANDRÉ SALAMIN¹

ABSTRACT

In longitudinal surveys, simple estimates of change, such as differences of percentages may not always be efficient enough to detect changes of practical relevance, especially in sub-populations. The use of models, which can represent the dependence structure of the longitudinal survey, can help to solve this problem. One of the main characteristics observed by the Swiss Labour Force Survey (SLFS) is the employment status. As the survey is designed as a rotating panel, the data from the SLFS are multivariate categorical data, where a large proportion of the response profiles are missing by design. The multivariate logistic model, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is attractive in this context, since it allows for dependent repeated observations and incomplete response profiles. We show that, using multivariate logistic regression, we can represent the complex dependence structure of the SLFS by a small number of parameters, and obtain more efficient estimates of change.

KEY WORDS: Longitudinal binary data; Multivariate logistic model; Labour force survey.

1. INTRODUCTION

One of the main objectives of the Swiss Labour Force Survey (SLFS), is to produce estimates of change for the percentages of the population in different employment statuses. Typically, simple estimates of change, such as the difference of the percentages of employed individuals between two years, are calculated for the whole population, and for a large number of sub-populations. In general, this is unsatisfactory, as the estimates for the sub-populations may not always be efficient enough to detect changes of practical relevance. The work presented here was motivated by the question, whether the use of models, which can represent the dependence structure of the survey, could help to solve this problem.

As the SLFS is designed as a rotating panel, we are dealing with longitudinal categorical data, for which a fairly large proportion of the response profiles, are incomplete by design. The focus of interest is on modelling marginal probabilities, namely, the probabilities to be in a given employment status, as a function of time and other covariates that define sub-populations. If the repeated observations of the employment status were independent, a natural approach would be to use logistic regression. The multivariate logistic model, introduced by Glonek and McCullagh (1995) as a generalisation of logistic regression, is attractive in this context, since it allows for dependent repeated observations and incomplete response profiles.

The aim of this paper is to show that, the ability of multivariate logistic regression to model the complex dependence structure of the SLFS data, leads to more efficient estimators of change. Although we illustrate the method using the SLFS data only, it is clearly of wider applicability.

There are a number of important issues that are not dealt with in this paper. As the SLFS data come from a complex survey, it can be argued that any analysis should take the sampling weights into account (Pfeffermann 1993). Here we use the unweighted data only. However, it can be shown, using the pseudo-likelihood approach of Binder (1983), that multivariate logistic regression can be extended to that situation (Salamín 1998). Non-response is always of great concern in sample surveys. Here, we consider only the incomplete response profiles that arise through the rotation of the panel, in which case, the hypothesis of missing completely at random, is reasonable. Note however, that multivariate logistic regression, is flexible enough to incorporate extra parameters for the incomplete profiles, arising from panel, attrition. Thus, the individuals which dropped out of the panel, could also have been included into the analysis. Finally, it is well known that classification errors may introduce large biases in the observed response profile probabilities, see *e.g.*, Pfeffermann, Skinner and Keith (1998). It would certainly be desirable to investigate how these biases affect the parameter estimates of multivariate logistic regression, which have interpretations in terms of marginal moments.

Log-linear models and marginal models are closely related to multivariate logistic regression, and are further discussed in Section 3. Here we discuss briefly transition models, random effects models, and survival analysis, in the context of the SLFS. Under a transition model, see *e.g.*, Diggle, Liang and Zeger (1994, Ch. 10) or Zeger and Liang (1992), the repeated observations of the employment status are correlated, because past employment statuses influence the present employment status. The focus of interest, are the transition probabilities between the different employment statuses, *e.g.*, the probability of being

¹ Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland.

employed, conditional on being unemployed in the past. In the regression setting, the past responses are treated as additional explanatory variables. An important issue, is the determination of the number of past responses to include as predictors. If the model for the transition probabilities is correctly specified, we can treat the repeated transitions for an individual as independent events, and use standard statistical methods, such as logistic regression. Under a random effect model, see *e.g.*, Diggle *et al.* (1994, Ch. 9), the probability of being in a given employment status, is a function of explanatory variables, where the regression coefficients vary from one individual to the next. This variability of the regression coefficients, reflects the natural heterogeneity of the individuals, due to unmeasured factors. Given the regression coefficients, the repeated observations of the employment status, are assumed to be independent. The correlation among the repeated observations, arises solely because we are unable to observe the true regression coefficients. This approach is most useful, when inference about individuals rather than population averages, is the focus of interest. In survival analysis, also called event history analysis in the econometric literature (Lancaster 1990), the focus is on modelling the transitions between employment statuses over time, as a function of explanatory variables. Here, the exact time at which a transition takes place, is important. In the SLFS, the employment status is observed once a year. The changes in employment status, that took place during the year preceding the interview, can be reconstructed. However, since this reconstruction is based on the self-assessment of the subjects, there may be some imprecision as regards prior status, and time of change of status. An analysis of the SLFS data based on this approach can be found in Gerfin (1996).

The article is organized as follows. We begin in Section 2 by describing the data, a subset of about 5000 individuals from the SLFS, which are used in the examples of Sections 4 and 5. In Section 3, we discuss multivariate logistic regression, and contrast it with the log-linear and marginal models. In Section 4, we illustrate the ability of multivariate logistic regression, to represent the complex dependence structure of the SLFS data, by a small number of parameters. In Section 5, we compare multivariate logistic regression with a simple estimator of change. It is shown that, using multivariate logistic regression, results in a gain in efficiency. Finally, we present in Section 6 our conclusions, and give directions for further work.

2. SWISS LABOUR FORCE SURVEY DATA

A detailed description of the sampling design and weighting procedure of the SLFS, can be found in Hulliger, Ries, Comment and Bender (1997). Here, we just recall some of the relevant aspects of this survey. The SLFS collects information on the employment of resident persons of age 15 or more in Switzerland. Starting in the second

quarter of 1991, a sample of about 16,000 persons are interviewed each year. The survey is designed as a rotating panel, with a time-in-sample of 5 years. During the start-up phase, *i.e.*, from 1992 to 1996, approximately one fifth of the original sample was rotated out each year, and replaced by a renewal sample. The units in the renewal samples then stayed in the panel for a full period of 5 years.

In the examples of Sections 4 and 5, we use the observations of the employment status, for the years 1992 to 1995, obtained from the individuals in the sample, of the canton of Vaud. The structure of the data, as well as the longitudinal and cross-sectional sample sizes, are shown in Table 1. Due to the sampling design, some of the response profiles are incomplete. For example, for the individuals that were selected in 1991 and rotated out of the sample in 1994, the period of observation, denoted (1)234, goes from 1991 to 1994. We use the notation (1)234, to emphasise the fact, that we do not use the observations taken in 1991.

Table 1
Structure of the Data, Longitudinal and Cross-sectional Sample Sizes Canton of Vaud, 1992-1995

First year in sample	Observation times for various parts of the sample				Period of observation	
91	92				(1)2	622
	92	93			(1)23	412
	92	93	94		(1)234	527
	92	93	94	95	(1)2345	481
92	92	93	94	95	2345	612
93		93	94	95	345	722
94			94	95	45	728
95				95	5	877
	2,654	2,754	3,070	3,420		4,981

Employment status is a nominal variable with three categories, defined as “employed”, “unemployed” and “out of the labour force”. In the examples of Section 4 and 5, we work with a binary variable, taking the value 1 if an individual is employed, and 2 if an individual is unemployed or out of the labour force. This is done solely to simplify the presentation of the multivariate logistic models. As the method can handle an arbitrary number of categories, it would be preferable, not to collapse the statuses in a real analysis. Caution must be exercised, if it is nevertheless necessary to combine some of the statuses, as heterogeneity of the statuses may introduce bias.

3. MULTIVARIATE LOGISTIC MODELS

The multivariate logistic model, introduced by Glonek and McCullagh (1995), can handle multivariate responses of either nominal or ordinal types, and either discrete or continuous explanatory variables. Here, we consider only multivariate binary responses and discrete predictors. The

multivariate logistic model, is an example of a generalized linear model, see McCullagh and Nelder (1989). Its link function, also called the multivariate logistic transformation, expresses the joint distribution of the response profiles, in terms of marginal moments of increasing order, the first two being marginal logits, and marginal log odds ratios. The link function has the property, termed reproducibility, that a multivariate logistic model, applies to any subset of the response vector. This property ensures that, the interpretations of the parameters are the same, regardless of the number of response variables, and whether or not higher order parameters are included. This makes multivariate logistic regression, especially attractive for the analysis of longitudinal data, where the repeated observations of an outcome arise on an equal footing, and where the number of repeated observations may vary from one individual to the next. Reproducibility is also the key to the ability of the model, to accommodate observations with incomplete responses. Note however, that we need to assume, that the data are missing completely at random, if the same parameters are to be used to model the complete and incomplete response profiles. The parameter estimates are found by maximum likelihood. A key step, is the inversion of the multivariate logistic transformation. For more than three responses, this may not always be possible, as there are then constraints among the parameters (Glonek and McCullagh 1995, Liang, Zeger and Qaqish 1992). Also, the presence of empty cells, may limit the order of the parameters that can be fitted.

The log-linear model is widely used to model multivariate binary data. In the saturated log-linear model, see *e.g.*, Liang *et al.* (1992), the canonical parameter associated with a subset of the variables, has an interpretation in terms of conditional probabilities given the rest of the variables, *e.g.*, the first and second order parameters are logits and log odds ratios, conditional on all the other responses. It follows that, the log-linear model is not reproducible, which makes it less preferable than multivariate logistic regression, for the analysis of longitudinal data. It is nevertheless possible, to build log-linear models that, as in the multivariate logistic model, have marginal logits as parameters. This leads to the marginal models (Diggle *et al.* 1994, Ch. 8). In these models, the dependence of the marginal probabilities on explanatory variables, is modelled separately from within-unit correlation. Under this approach, the parameters are not estimated by maximum likelihood. Rather, only the structure of the correlation, between the repeated observations of an outcome is specified, and the parameters are estimated by solving generalized estimating equations (GEE), a multivariate analogue of quasi-likelihood (McCullagh and Nelder 1989). A number of specifications of the correlation structure have been proposed, for example Liang *et al.* (1992) use the marginal log odds ratios, as in Glonek and McCullagh (1995). We have made some comparisons between multivariate logistic regression and PROC GENMOD of SAS (release 6.12).

This procedure has the ability to fit correlated response models by the GEE method. We found very similar estimates of the marginal logits. The GEE method appeared to be slightly less efficient than multivariate logistic regression. A limitation of the GEE method is that, it cannot yield estimates of the response profile probabilities, but only of the marginal probabilities. By contrast, the multivariate logistic model does not have this limitation, since its parameters are estimated by maximum likelihood.

Following Glonek and McCullagh (1995), we discuss in Section 3.1 the multivariate logistic transformation, and we give, in Section 3.2, the algorithm for maximum likelihood.

3.1 Multivariate Logistic Transformation

Let Y_1, Y_2, \dots, Y_d be d repeated observations, taken at times $t_1 < t_2 < \dots < t_d$, of the same binary variable, and let

$$\pi_{i_1 i_2 \dots i_d} = P(Y_1 = i_1, Y_2 = i_2, \dots, Y_d = i_d),$$

where i_1, i_2, \dots, i_d are all either 1 or 2, be the joint probabilities of the random variables Y_1, Y_2, \dots, Y_d . In the multivariate logistic model, the joint probabilities of Y_1, Y_2, \dots, Y_d are parameterized in terms of marginal logits, marginal log odds ratios, and contrasts of marginal log odds ratios. This parameterization can be written as $\eta = C^T \log(L\pi)$, where π is the vector of dimension $q = 2^d$

$$\pi = (\pi_{11\dots 11}, \pi_{11\dots 12}, \dots, \pi_{22\dots 21}, \pi_{22\dots 22})^T,$$

and where, the matrices L and C are tensor products of suitably chosen marginal indicator and contrast matrices respectively. The matrices L and C , which depend on the length d of the observation period, are defined recursively, beginning with $L_0 = C_0 = 1$, as

$$L_d = \begin{bmatrix} L_{d-1} \otimes 1_2^T \\ L_{d-1} \otimes \tilde{L} \end{bmatrix}$$

and

$$C_d = \begin{bmatrix} C_{d-1} & 0 \\ 0 & C_{d-1} \otimes \tilde{C} \end{bmatrix},$$

where $1_2^T = (1, 1)$, \tilde{L} is the two by two identity matrix, and $\tilde{C} = (1, -1)^T$ (Glonek and McCullagh 1994).

To illustrate matters, we consider periods of observation of length $d = 1, 2, 3, 4$. For $d = 1$, $\pi = (\pi_1, \pi_2)^T$ and $\eta = (\eta_0, \eta_1)^T = (\log \pi_+, \text{logit } Y_1)^T$, where the plus subscript indicates summation, and $\text{logit } Y_1$ is defined as

$$\text{logit } Y_1 = \log \frac{P(Y_1 = 1)}{P(Y_1 = 2)} = \log \frac{\pi_1}{1 - \pi_1} = \log \frac{\pi_1}{\pi_2}.$$

In that case the multivariate logistic transformation is equivalent to the usual logistic transformation. Note that, although the parameter $\eta_0 = \log \pi_+ = 0$ is strictly superfluous, it is convenient to retain it, as a means of ensuring that the mapping $\pi \rightarrow \eta$ is of full rank, and also expressing the requirement that $\pi_+ = 1$.

For $d = 2$, $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12})^T =$$

$$(\log \pi_{++}, \log \text{it } Y_1, \log \text{it } Y_2, \log OR(Y_1, Y_2))^T$$

where

$$OR(Y_1, Y_2) =$$

$$\frac{P(Y_1 = 1, Y_2 = 1) P(Y_1 = 2, Y_2 = 2)}{P(Y_1 = 1, Y_2 = 2) P(Y_1 = 2, Y_2 = 1)} = \frac{\pi_{11} \pi_{22}}{\pi_{12} \pi_{21}}$$

is the odds ratio, a quantity measuring the association between the variables Y_1 and Y_2 . The parameters η_1 and η_2 are the marginal logits at times t_1 and t_2 , for example

$$\eta_1 = \log \text{it } Y_1 = \log \frac{\pi_{1+}}{(1 - \pi_{1+})}.$$

For $d = 3$, $\pi = (\pi_{111}, \pi_{112}, \dots, \pi_{221}, \pi_{222})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123})^T.$$

The parameters η_1, η_2 and η_3 are the marginal logits at times t_1, t_2 and t_3 . The parameters η_{12}, η_{13} and η_{23} are the log odds ratios of the corresponding two-dimensional marginal tables, for example

$$\eta_{23} = \log OR(Y_2, Y_3) = \log \frac{\pi_{+11} \pi_{+22}}{\pi_{+12} \pi_{+21}}.$$

The parameter η_{123} is a contrast of log odds ratios given by

$$\begin{aligned} \eta_{123} &= \log OR(Y_1, Y_2 | Y_3 = 1) - \log OR(Y_1, Y_2 | Y_3 = 2) \\ &= \log \frac{\pi_{111} \pi_{221}}{\pi_{121} \pi_{211}} - \log \frac{\pi_{112} \pi_{222}}{\pi_{122} \pi_{212}}. \end{aligned}$$

For $d = 4$, $\pi = (\pi_{1111}, \pi_{1112}, \dots, \pi_{2221}, \pi_{2222})^T$ and

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123}, \eta_4, \eta_{14}, \eta_{24}, \eta_{124}, \eta_{34}, \eta_{134}, \eta_{234}, \eta_{1234})^T.$$

The parameters η_i, η_{ij} and η_{ijk} , where $1 \leq i < j < k \leq 4$, are defined as above, using the appropriate marginal tables. The parameter η_{1234} is a contrast of log odds ratios given by

$$\eta_{1234} = \log OR(Y_1, Y_2 | Y_3 = 1, Y_4 = 1)$$

$$- \log OR(Y_1, Y_2 | Y_3 = 1, Y_4 = 2)$$

$$- \log OR(Y_1, Y_2 | Y_3 = 2, Y_4 = 1)$$

$$+ \log OR(Y_1, Y_2 | Y_3 = 2, Y_4 = 2).$$

A key step in maximum likelihood estimation is the computation of the inverse of the multivariate logistic transformation. To ensure that $\pi > 0$, we work with $\pi = \exp v$, *i.e.*, we seek to solve for v in the equation $\eta = C^T \log(L \exp v)$. In general, no explicit solution is available, so an iterative method must be used. In particular, the Newton-Raphson iterations can be applied as described below. For clarity, we define the two functions $\phi(\pi) = C^T \log(L \pi)$ and $\psi(v) = \phi(\exp v)$.

- (i) Begin with an initial approximation v_0 .
- (ii) Then take $v_{k+1} = v_k - [D\psi(v_k)]^{-1} (\phi(\exp v_k) - \eta)$, where $D\psi(v)$ is the Jacobian matrix of the function $\psi(v)$, and iterate until convergence.

The Jacobian matrices of the function $\phi(\pi)$ and $\psi(v)$ are given respectively by $D\phi(\pi) = C^T (\text{diag } L \pi)^{-1} L$ and $D\psi(v) = D\phi(\exp v) \cdot \text{diag}(\exp v)$.

3.2 Maximum Likelihood Estimation

For a binary response variable observed at d time points, there are $q = 2^d$ possible response profiles $i = (i_1, \dots, i_d)$, where i_1, i_2, \dots, i_d are all either 1 or 2. For each profile $i = (i_1, \dots, i_d)$, we define the indicator variable $Y_{i_1 \dots i_d}$, which is equal to 1 if the profile i has been observed, and 0 otherwise. We then have

$$P(Y_{i_1 \dots i_d} = 1) = P(Y_1 = i_1, \dots, Y_d = i_d) = \pi_{i_1 \dots i_d}.$$

Defining the q -dimensional vectors

$$Y = (Y_{11\dots 11}, Y_{11\dots 12}, \dots, Y_{22\dots 21}, Y_{22\dots 22})^T$$

and

$$\pi = (\pi_{11\dots 11}, \pi_{11\dots 12}, \dots, \pi_{22\dots 21}, \pi_{22\dots 22})^T,$$

we may then write $Y \sim M(1, \pi)$, *i.e.*, Y is a multinomial vector with $q = 2^d$ categories, whose probabilities are given by the vector π .

The multivariate logistic regression models, are then defined to be those of the form $\eta = X\beta$ where X is a $q \times p$ matrix of explanatory variables, β is a p -dimensional vector of unknown parameters, and $\eta = C^T \log(L \pi) = \phi(\pi)$.

If we let y be one observation of the random vector Y , then we may write the kernel of the log likelihood function as $l(\beta; y) = y^T \log \pi(\beta)$ where, using the inverse of the

multivariate logistic transformation, we can express the joint probabilities π as a function of the unknown parameter β , as $\pi(\beta) = \varphi^{-1}(X\beta)$. The score vector is given by

$$s(\beta) = s(\beta, y, X) = D\pi(\beta)^T (\text{diag} \pi(\beta))^{-1} y,$$

where $D\pi(\beta)$, the Jacobian matrix of the function $\pi(\beta)$, relating the parameter β to the vector of probabilities π , is given by $D\pi(\beta) = [D\varphi(\varphi^{-1}(X\beta))]^{-1} X$, and where $D\varphi(\pi) = C^T (\text{diag} L\pi)^{-1} L$, is the Jacobian matrix of the link function. The information matrix is defined as $\mathfrak{S}(\beta) = Es(\beta)s(\beta)^T$. Now it follows from the assumption on the distribution of Y , that $E(YY^T) = \text{diag} \pi$, from which we may deduce that

$$\mathfrak{S}(\beta) = \mathfrak{S}(\beta, X) = D\pi(\beta)^T (\text{diag} \pi(\beta))^{-1} D\pi(\beta).$$

If we have n independent observations $y_k \sim M(1, \pi_k)$, $k = 1, \dots, n$, where $\eta_k = C^T \log(L\pi_k) = X_k \beta$, then the score vector and the information matrix are given by $s(\beta) = \sum_{k=1}^n s(\beta, y_k, X_k)$ and $\mathfrak{S}(\beta) = \sum_{k=1}^n \mathfrak{S}(\beta, X_k)$.

The maximum likelihood estimator of β is the solution of $s(\beta) = 0$, that can be found by using the Fisher scoring algorithm which, starting from some initial value β_0 , iterates the sequence $\beta_{m+1} = \beta_m + \mathfrak{S}_m^{-1}(\beta_m) s(\beta_m)$ until convergence.

Incomplete response profiles can readily be incorporated into the analysis. In particular, if some subset of the response variables Y_1, Y_2, \dots, Y_c is recorded for a particular unit, then the probability distribution on that c -dimensional marginal table is multinomial, and, as a consequence of the reproducibility of the multivariate logistic transformation, a multivariate logistic regression model applies to the table of probabilities. Furthermore, the design matrix relating the marginal probabilities to β , is constructed by selecting the appropriate rows of the full design matrix, that would be used if complete data were available for that unit.

4. MODELS FOR LONGITUDINAL DEPENDENCE

In this section we illustrate, using the SLFS data of Section 2, how multivariate logistic regression can be applied to describe the dependence between the repeated observations of the employment status. We do not intend to carry out an exhaustive search for a best model, but rather to demonstrate the ability of the method, to represent a complex dependence structure by a small number of parameters.

We consider 6 models of decreasing complexity, see Table 2. For all 6 models, we have one parameter for each of the marginal logits corresponding to a given observation time. Symbolically, this is denoted by $\eta_i = \beta_i$. Since the observation times are the 2nd quarter of the years 1992 to 1995, we take $i = 2, 3, 4, 5$. Thus β_3 , say, corresponds to the logit of the probability of being employed in 1993.

Similarly, the indices for the higher order parameters run from 2 to 5. For model 1 we take a saturated model for the longitudinal dependence, *i.e.*, we have one parameter for each of the interactions of order 2, 3 or 4 within each period of observation. For the models 2 to 5, we assume that the interactions of order 3 and 4, are all equal to zero. The longitudinal dependence is then described in terms of log odds ratios only. For model 2, we take a saturated model for the log odds ratios. In model 3 we drop the covariate period of observation, *i.e.*, we suppose that the log odds ratios are the same for all the periods of observation. In model 4, we use stationary log odds ratios, *i.e.*, log odds ratios which depend only on the difference between times of observation. Note that the parameter γ_1 in model 4, corresponds to the constraint $\beta_{23} = \beta_{34} = \beta_{45}$ on the parameters of model 3, and similarly for γ_2 and γ_3 . In model 5, a linear model for the stationary log odds ratios is assumed. In model 6, finally, we assume that the observations taken at different times, are independent. Note that, in that case, multivariate logistic regression is equivalent to ordinary logistic regression.

Table 2
Six Models for Longitudinal Dependence

Model	Parameters		
	Marginal logits	Log odds ratios	3 rd and 4 th order parameters
1	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij, \text{period}}$	$\eta_{ijk} = \beta_{ijk, \text{period}}, \eta_{ijkl} = \beta_{ijkl, \text{period}}$
2	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij, \text{period}}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
3	$\eta_i = \beta_i$	$\eta_{ij} = \beta_{ij}$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
4	$\eta_i = \beta_i$	$\eta_{ij} = \gamma_{ i-j }$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
5	$\eta_i = \beta_i$	$\eta_{ij} = \delta + \gamma \cdot i-j $	$\eta_{ijk} = 0, \eta_{ijkl} = 0$
6	$\eta_i = \beta_i$	$\eta_{ij} = 0$	$\eta_{ijk} = 0, \eta_{ijkl} = 0$

The parameter estimates for the models 2 to 6, are given in Table 3. The number of parameters and the values of the log likelihood function at the maximum likelihood estimates, can be found in Table 4 where, for comparison, we also included the log likelihood for the fully saturated model.

Overall, we notice that the assumed form of the longitudinal dependence, appears to have little effect on the estimates of the marginal logits. This is a desirable property, as the marginal logits would typically be the parameters of interest. The standard errors of the marginal logits, are almost the same for the models that take into account the longitudinal dependence, but are inflated by about 15% for ordinary logistic regression (model 6). It can also be shown that the estimates of the marginal logits are positively correlated under the models that assume a longitudinal dependence, and uncorrelated for ordinary logistic regression. For the example considered here, the

correlation was found to lie between 0.4 and 0.8. Thus, modelling the longitudinal dependence, leads also to more efficient estimates of the difference of marginal logits.

It can be seen from the fit of model 1, that the interaction parameters of order 3 and 4, are not significantly different from 0. This suggests that the longitudinal dependence can be described by the log odds ratios only. This hypothesis is corroborated by the incremental deviance of model 2 with respect to model 1, which is found to be 7.9, on 12 degrees of freedom. Further, all the parameters of model 2 are significantly different from 0, and an examination of the standardised residuals for the fitted probabilities of the response profiles, does not reveal any anomaly. For applications in official statistics, model 2 would be the preferred model, since it is based on as few assumptions as possible, while still allowing a substantial reduction in the number of parameters, thus rendering less acute the danger

of sparse tables when longer periods of observation and models with more covariates are considered.

The models 3, 4 and 5 show that, it would nevertheless be possible to greatly simplify the description of the longitudinal dependence, without losing too much information. In going from model 2 to model 5, we observe that the deviance from the fully saturated model, does not increase much, see Table 4. Further, an examination of the residuals shows that, the models 3, 4 and 5 fit the data almost as well as model 2. On the other hand, while model 2 requires 20 parameters to describe the longitudinal dependence, model 5 needs only 2 parameters. This must be contrasted with model 6, which assumes independence between observations taken at different times: the log likelihood is much smaller than for the fully saturated model, see Table 4, and the fit to the data is poor.

Table 3
Parameter Estimates and Standard Errors

Parameter	Period	Model 2	Model 3	Model 4	Model 5	Model 6
logit 92		0.6348 (0.0350)	0.6360 (0.0352)	0.6348 (0.0352)	0.6347 (0.0352)	0.6471 (0.0409)
logit 93		0.5555 (0.0335)	0.5570 (0.0338)	0.5597 (0.0335)	0.5601 (0.0335)	0.5509 (0.0396)
logit 94		0.5440 (0.0324)	0.5407 (0.0325)	0.5402 (0.0326)	0.5397 (0.0325)	0.5377 (0.0374)
logit 95		0.4699 (0.0317)	0.4711 (0.0320)	0.4710 (0.0320)	0.4712 (0.0320)	0.4705 (0.0351)
β_{23}	(1)23	4.2563 (0.3311)	4.2579 (0.1465)			
	(1)234	4.2003 (0.2894)				
	(1)2345	4.0859 (0.2954)				
	2345	4.4830 (0.2841)				
β_{34}	(1)234	4.0894 (0.2794)	4.1111 (0.1310)			
	(1)2345	3.9611 (0.2840)				
	2345	4.0989 (0.2600)				
	345	4.2490 (0.2468)				
β_{45}	(1)2345	5.3992 (0.3854)	4.5561 (0.1389)			
	2345	3.9779 (0.2544)				
	345	4.7288 (0.2735)				
	45	4.5069 (0.2600)				
β_{24}	(1)234	3.7168 (0.2641)	3.8371 (0.1442)			
	(1)2345	4.2560 (0.3059)				
	2345	3.5330 (0.2370)				
β_{35}	(1)2345	4.4000 (0.3098)	3.7913 (0.1334)			
	2345	3.6493 (0.2396)				
	345	3.6116 (0.2192)				
β_{25}	(1)2345	4.3984 (0.3173)	3.5774 (0.1530)			
	2345	3.2209 (0.2256)				
γ_1				4.3260 (0.0928)		
γ_2				3.8519 (0.1050)		
γ_3				3.5340 (0.1495)		
δ					4.7341 (0.1266)	
γ					-0.4191 (0.0653)	

Table 4
Number of Parameters and Value of the Log Likelihood
Function at the Maximum Likelihood Estimates

Model	Number of parameters of order					Log likelihood
	1	2	3	4	Total	
Full Model	20	20	10	2	52	-5342.7
1	4	20	10	2	36	-5345.4
2	4	20	0	0	24	-5349.4
3	4	6	0	0	10	-5365.2
4	4	3	0	0	7	-5368.9
5	4	2	0	0	6	-5369.5
6	4	0	0	0	4	-7815.3

5. COMPARISON WITH SIMPLE ESTIMATE OF CHANGE

In this section we concentrate on the estimation of the difference of the probabilities of being employed between any two given years. We show that, estimates based on multivariate logistic regression, are more efficient than simple estimates defined as the difference of the proportions of employed individuals.

The model considered here, is the model 2 of Section 4, with sex as an additional explanatory variable. We have, for each sex, one parameter for each of the marginal logits corresponding to a given year. The longitudinal dependence is accounted for by a saturated model for the log odds ratios. The third and fourth order parameters are set to 0. This model has therefore 8 parameters for the marginal logits, and 40 parameters for the log odds ratios: 2 sexes \times 20 odds ratios within periods of observation, see Table 3. By inverting the multivariate logistic transformation, estimates of the probability of being employed, and of their differences between any two given years, can also be computed.

A simple estimator of change is given by the difference of the proportions of employed individuals between any two given years. Its variance, which takes into account the overlap of the two samples, is given by

$$\frac{1}{n+r} \pi_{1+}(1 - \pi_{1+}) + \frac{1}{n+c} \pi_{+1}(1 - \pi_{+1}) - 2 \frac{n}{(n+r)(n+c)} (\pi_{11} - \pi_{1+} \pi_{+1}),$$

where n is the number of cases for which observations are available for both years, r and c are the number of cases for which observations are available for only one year, π_{11} is the probability of being employed during both years, and π_{1+} and π_{+1} are the marginal probabilities of being employed.

Tables 5 shows, for the SLFS data of Section 2, the estimates of the difference of the probability of being

employed under both methods. Note that both methods yield similar estimates of change. The standard errors of the simple estimates, are on the average, 30% larger than for multivariate logistic regression. The corresponding mean relative efficiency of multivariate logistic regression, with respect to the simple estimates, is 1.7. By comparison, the mean relative efficiency of multivariate logistic regression with respect to ordinary logistic regression, is 3.2.

Table 5
Change in the Probability of Being Employed
Canton of Vaud, 1992-1995

Comparison		Multivariate logistic regression	Simple estimate
Woman	92 vs. 93	0.0138 (0.0090)	0.0136 (0.0115)
	92 vs. 94	0.0184 (0.0102)	0.0168 (0.0134)
	92 vs. 95	0.0375 (0.0109)	0.0356 (0.0149)
	93 vs. 94	0.0047 (0.0087)	0.0031 (0.0107)
	93 vs. 95	0.0238 (0.0095)	0.0219 (0.0128)
	94 vs. 95	0.0191 (0.0076)	0.0188 (0.0100)
Men	92 vs. 93	0.0220 (0.0095)	0.0283 (0.0116)
	92 vs. 94	0.0245 (0.0102)	0.0334 (0.0133)
	92 vs. 95	0.0387 (0.0106)	0.0452 (0.0144)
	93 vs. 94	0.0024 (0.0092)	0.0052 (0.0111)
	93 vs. 95	0.0167 (0.0098)	0.0169 (0.0130)
	94 vs. 95	0.0143 (0.0080)	0.0117 (0.0102)

6. CONCLUSIONS

The analyses of the SLFS data presented here, have shown the usefulness of multivariate logistic regression. Modelling the longitudinal dependence is necessary, in order to obtain a satisfactory fit of the observed response profile probabilities. Ignoring the longitudinal dependence, we still obtain acceptable point estimates of the marginal logits, but the information on the detailed structure of the data is lost. Modelling the longitudinal dependence leads also to more efficient estimates of the marginal parameters and of change, when compared with ordinary logistic regression, and a simple estimator of change. Finally, the ability of multivariate logistic regression to represent a complex dependence structure, by a small number of parameters, has also been illustrated.

Using the results of Glonek and McCullagh (1995), it is possible to extend the examples presented here, to multivariate responses of either nominal or ordinal types, with either discrete or continuous explanatory variables. The method can also be extended, to take the sampling weights into account (Salamin 1998). For the SLFS, it was found that the sampling weights have little effect on the parameter estimates of the multivariate logistic model. The standard error of the parameter estimates, was inflated by about 15%. This moderate increase of the variability of the parameter estimates due to the sampling weights, is plausible.

Indeed, as in the SLFS, only one person per household is selected, a large cluster effect was not expected.

Apart from the sort of analyses presented here, multivariate logistic regression may also be used for modelling non-response probabilities in longitudinal studies. Such models may be useful when the sampling weights need to be adjusted for non-response. The ability of multivariate logistic regression to give a parsimonious model of the data, may also be of interest in small-area estimation. In particular, estimators for a given geographical region could be based on models for an appropriately chosen larger region.

Although we did not encounter serious problems in the examples presented here, further work may need to be done on the problem of sparse tables. A critical step, when there are a large number of empty cells, is the inversion of the multivariate logistic transformation. The approach of Lang (1996), where the inversion of the link function is avoided, by specifying the models through constraints, may be of interest in this context. Another area of investigation is the influence of the classification errors on the parameter estimates of the multivariate logistic model.

ACKNOWLEDGEMENTS

This paper benefited from discussions with colleagues at the Swiss Federal Statistical Office, among which Beat Hulliger and Philippe Eichenberger deserve special mention. The help of Ariane Bender, from the Section responsible for the Swiss Labour Force Survey, in preparing and understanding the data, was extremely valuable. The author also thanks the editor and two referees for their excellent suggestions that have led to a significant improvement of this paper.

REFERENCES

- BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- DIGGLE, P.J., LIANG, K.-Y., and ZEGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.
- GERFIN, M. (1996). *Entwicklung von ökonometrischen Modellen zur Analyse der Dynamik auf dem schweizerischen Arbeitsmarkt*. SLFS-News, Swiss Federal Statistical Office, Berne.
- GLONEK, G.F.V., and McCULLAGH, P. (1994). Multivariate Logistic Models. Technical Report 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide.
- GLONEK, G.F.V., and McCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society, B*, 57, 533-546.
- HULLIGER, B., RIES, A., COMMENT, T., and BENDER, A. (1997). Weighting the Swiss Labour Force Survey. (Eds. C. Malaguerra, S. Morgenthaler and E. Ronchetti). In *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switzerland*, Basel: Birkhäuser Verlag.
- LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.
- LANG, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, 24, 726-752.
- LIANG, K.-Y., ZEGER, S.L., and QAQISH, B. (1992). Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society, B*, 54, 3-40.
- McCULLAGH, P., and NELDER, J.A. (1989). *Generalized Linear Models*, (2nd edn.). London: Chapman and Hall.
- PFEFFERMANN, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.
- PFEFFERMANN, D., SKINNER, C., and KEITH, H. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, A*, 161, Part 1, 13-32.
- SALAMIN, P.-A. (1998). Multivariate logistic regression for data from complex surveys. To appear *Proceedings: Symposium '98, Longitudinal Analysis for Complex Surveys*, Statistics Canada, May 1998.
- ZEGER, S.L., and LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.

Price Index Surveys as Quasi-Longitudinal Studies

ALAN H. DORFMAN¹

ABSTRACT

To calculate price indexes, data on "the same item" (actually a collection of items narrowly defined) must be collected across time periods. The question arises whether such "quasi-longitudinal" data can be modeled in such a way as to shed light on what a price index is. Leading thinkers on price indexes have questioned the feasibility of using statistical modeling at all for characterizing price indexes. This paper suggests a simple state space model of price data, yielding a consumer price index that is given in terms of the parameters of the model.

KEY WORDS: Random walk plus noise model; State space model; Laspeyres index; Paasche index; Geometric price index.

1. INTRODUCTION

Survey sampling for calculation of a consumer price index is characterized by following a given item across time to determine its prices at a succession of times. Only it is not, typically, exactly the same item that is followed – it is not this particular can of Brand Y Tomato Soup at Outlet Z the price of which is repeatedly ascertained, for this particular can is likely to have been sold and consumed, by the time of the next visit of the survey sampler – but rather a succession of items, each fitting the same description ("Brand Y 8 oz. Can of Tomato Soup with Herring, sold at Outlet Z"), the price of which is collected at different times. In other words, it is essentially a group of items fitting a narrow description which is followed across time. For this reason consumer price index surveys may be termed "quasi-longitudinal" as opposed to longitudinal surveys, which would follow individual items across time. Nonetheless, it is reasonable to hope that, having repeated measurements across time might lead to estimation procedures which could capitalize on the time series aspect of such surveys.

In the light of that hope, this paper considers a question which has by and large been ignored by statisticians and economists, or, when not ignored, been answered in the negative: Can a consumer price index (CPI) be treated from a statistical point of view? That is, can the parameter, which characterizes the "change in the cost of living" from one period to another, and which price index surveys attempt to estimate, be defined in terms of a stochastic model?

Aldrich (1992) gives an historic interpretation of early attempts by Jevons and especially Edgeworth, to incorporate distributional assumptions into the CPI. Recent papers on stochastic modeling of the CPI, are those by Balk (1980), Clements and Izan (1981, 1987), Bryan and Cecchetti (1993), Kott (1984) and Selvanathan and Rao (1994). Diewert (1995) reviews and criticizes these attempts, taking an argument of Keynes (1930) as decisive grounds for rejecting the stochastic approach.

In this paper, a specific approach to modeling the price index using state space models is suggested, and a specific state space model tentatively suggested. This model is applied to scanner data to demonstrate the feasibility of an index based on it. The approach we contemplate, circumvents the Keynesian criticism in fundamental ways, and offers the prospect of the many advantages that sound statistical modeling can bring, including, possibly, simplifications of the survey sampling process.

In what follows, we first briefly review the definition of a price index, and the two (non-stochastic approaches) which have dominated consideration of choice of index (Section 2). We review the Bryan and Cecchetti (1993) example of a statistical model for the price index, and Diewert's formulation of Keynes' objection (Section 3). We then introduce an approach to modeling a consumer price index, that circumvents the Keynes-Diewert difficulties, and that leads naturally to the use of state space models (Section 4). We present results of applying a relatively simple random walk plus noise model to scanner data from the A.C. Nielsen Academic Data Base (section 5). We assess the new index in Section 6, mentioning further research that might be useful.

2. BACKGROUND

What is meant by a Consumer Price Index (CPI) is a single number indicating how the purchasing power of the consumer has changed from one period t' to another t . Its raw ingredients consist of prices for the variety of available items at (at least) the two time periods

$$p_{\tau} = (p_{\tau 1}, \dots, p_{\tau N}), \tau = t', t$$

as well as quantities of the items sold

$$q_{\tau} = (q_{\tau 1}, \dots, q_{\tau N}), \tau = t', t.$$

¹ Alan H. Dorfman, U.S. Bureau of Labor Statistics, Room 4915, 2 Massachusetts Ave. N.E., Washington, D.C., 20212-0001, U.S.A.; e-mail: dorfman_a@bls.gov.

(Often however in practice quantity data from the periods in question are unavailable, and one makes do with some form of surrogate.) The CPI is derived from a “formula” that uses these raw ingredients:

$$I_{t't} = f(p_t, p_{t'}, q_t, q_{t'}),$$

where $f(\cdot)$ is a function of one of many possible forms. Most such forms have a long history, and have been extensively discussed in the index literature.

As examples, we mention here the Laspeyres index

$$L_{t't} = \frac{\sum_{i=1}^N q_{t'i} P_{ti}}{\sum_{i=1}^N q_{t'i} P_{t'i}} = \sum_{i=1}^N f_{t'i} r_{t'i},$$

with $f_{t'i} = q_{t'i} P_{ti} / \sum_{i=1}^N q_{t'i} P_{t'i}$ the “relative expenditures”, and $r_{t'i} = p_{ti} / p_{t'i}$ the “price relatives”. The Laspeyres index uses the quantities from the earlier time period, as a fixed basis of comparison of the earlier and later prices. The Laspeyres index (or a close variant) has tended to be the index most targeted by governments, because of its simplicity and intelligibility to the layperson.

The natural counterpart to the Laspeyres is the Paasche index

$$P_{t't} = \frac{\sum_{i=1}^N q_{ti} P_{ti}}{\sum_{i=1}^N q_{ti} P_{t'i}}$$

which standardizes the prices by the later period quantities. Most indices following other formulas will tend to fall between the Paasche and Laspeyres.

For later reference in this paper, we mention an index based on the geometric mean, with fixed non-negative weights f_i , adding to 1:

$$G_{t't} = \prod_{i=1}^N \left(\frac{p_{ti}}{p_{t'i}} \right)^{f_i}.$$

This is sometimes referred to as the “Geomean”.

Fisher (1922) discusses these and many other index formulae. He introduces what has come to be called the “Test Approach”, for deciding among the variety of candidates for the formula $f(\cdot)$: this approach lays out properties (“tests”), which a reasonable index would seem to require, and then examines to what extent each index formula satisfies them.

One of the tests is the Time Reversal Test: $I_{t't} I_{t't'} = 1$. Two indices which continue to exercise their sway in the

world, but fail this test are, the Carli-Sauerbach index $C_{t't} = \sum_{i=1}^N f_i p_{ti} / p_{t'i}$ and a geomean $\tilde{G}_{t't} = \prod_{i=1}^N (p_{ti} / p_{t'i})^{f_i}$ which employs first period expenditures instead of fixed weights. One readily shows that $C_{t't} C_{t't'} \geq 1$, using the Cauchy-Schwartz inequality, suggesting that this index will run too high.

If an increase in prices on item i tends to give an increase in expenditure share, then $\tilde{G}_{t't} \tilde{G}_{t't'} \leq 1$, so that under such conditions, the first-period-geomean tends to run too low. If an increase in prices on item i tends to give a decrease in expenditure share, then $\tilde{G}_{t't}$ runs too high. In general, we can expect this to be a rather erratic index.

This suggests the following maxim: price indices of the form of a geometric mean, should not have weights tied to prices at one of the periods being compared; those of the form of an arithmetic mean should not have weights independent of those prices.

By contrast with $\tilde{G}_{t't}$, the geomean $G_{t't} = \prod_{i=1}^N (p_{ti} / p_{t'i})^{f_i}$ which has fixed weights, is the unique index which satisfies the five axioms on price indices in Balk (1995), and the “circularity test”, which says that, for $t' < t^* < t$, $I_{t't} = I_{t't^*} I_{t^*t}$. Time reversal is an immediate consequence.

Indices which pass most of the tests, tend to be ones incorporating quantity information from both periods; for example, the Fisher index

$$F_{t't} = (L_{t't} P_{t't})^{1/2}$$

and the Törnqvist index

$$T_{t't} = \prod_{i=1}^N \left(\frac{p_{ti}}{p_{t'i}} \right)^{f_{ti}},$$

with $f_{ti} = (f_{t'i} + f_{ti})/2$. The Fisher and Törnqvist are frequently practically indistinguishable. Further discussion of the test approach, may be found in Balk (1995), Diewert (1987), and Eichhorn and Voeller (1976).

The second approach to assessing index formulas is the “economic” approach. This defines a generic index of the form

$$I_{t't} = \frac{C(p_t, U)}{C(p_{t'}, U)},$$

where $U = U(q_1, \dots, q_N)$ is a well-defined “utility function”, and $C(p_t, U)$ is the minimal cost at prices p_t , of achieving the standard of living, or “utility” U . For a particular utility function U , one inquires whether a particular formula can be regarded as a good approximation to the corresponding cost of living index. Like the test approach, this tends to yield indexes incorporating quantity information from both periods. See Diewert (1987) for further elaboration.

3. THE STOCHASTIC APPROACH

Aldrich (1992) gives the early history of attempts to model price relatives or logarithms of price relatives, using a common parameter that represents the overall rate of growth in prices. A basic theme of his paper is, that the stochastic approach to price indices, while being an early example of the application of statistics to economic concerns, died a natural death. Diewert (1995) also discusses these, as well as more recent examples of the statistical modeling of price relatives. The difficulty which, following Keynes (1930), Diewert finds with such use of models is exemplified by a model of Clements and Izan (1987).

The period from t' to t is divided into equi-temporal pieces, giving relatively short intervals generically represented as being from $t-1$ to t . The logarithm of the price relatives for such a "micro-period", is given by

$$\log \left(\frac{p_{ti}}{p_{t-1,i}} \right) = \pi_t + \beta_i + \varepsilon_{ti}, \quad (1)$$

with $\varepsilon_{ti} \sim (0, \sigma_i^2/f_i)$. In their model, the f_i 's are the average expenditure share of the i -th item over the period t' to t . For the sake of identifiability, it is assumed that $\sum_{i=1}^N f_i \beta_i = 0$. These assumptions lead to a maximum likelihood estimator

$$\hat{\pi}_t = \sum_{i=1}^N f_i \log \left(\frac{p_{ti}}{p_{t'i}} \right),$$

giving an MLE of the price short period price trend as

$$\exp(\hat{\pi}_t) = \prod_{i=1}^N \left(\frac{p_{ti}}{p_{t'i}} \right)^{f_i};$$

that is, based on their stochastic model, one derives a geometric index, with weights f_i , akin to that for the Törnqvist.

Estimates of the β_i and of σ^2 can also be derived, as well as estimates of precision, for example, of the variance of $\hat{\pi}$. Thus, a new statistical foundation seems to be put under an old estimator.

Diewert (1995) raises several objections, none of which can be taken lightly. The chief of these is

"... the fundamental objection of Keynes (Keynes 1930, p. 78): 'The hypothetical change in the price level [$\exp(\pi_t)$] which should have occurred if there had been no changes in relative prices, is no longer relevant if relative prices have in fact changed – for the change in relative prices has in itself affected the price level'."

If, say, the price of bread relative to the price of automobiles changes, then by that very fact, the overall price level changes.

Keynes' objection is not entirely clear. Why can't there be two aspects of price change, one overall, and the other particular? However, it is not hard to agree that the individual price trends are primary; an overall price trend can only be some weighted sum of these. Diewert offers the following translation into terms of a model, of Keynes' objection. Since we must have the overall price trend of the form

$$\pi_t^* = \sum_{i=1}^N f_i \beta_{ti}^*,$$

the model (1) needs to be replaced by

$$\log \left(\frac{p_{ti}}{p_{t-1,i}} \right) = \pi_t + \beta_{ti} + \varepsilon_{ti}, \quad (2)$$

with $\beta_{ti} = \pi_t - \beta_{ti}^*$ and $\sum_{i=1}^N f_i \beta_{ti} = 0$. The crucial difference between this and (1) is that now the item parameters β_{ti} are indexed by time. But "then the resulting model has too many parameters to be identified." This would suffice to nullify the approach.

Diewert (1995) does not discuss the much more complicated time-series model of Bryan and Cecchetti (1993). Of preceding papers, it is probably the closest to our present paper, involving a complicated state space model and use of the Kalman Filter. Like the other papers Diewert reviews, it is subject to Keynes' objection.

4. PRICE INDEXES RECONSIDERED

4.1 Common Presuppositions

The stochastic modeling of price behavior given in the last section, whether embodied in equation (1) or (2), or some similar model, has three notable characteristics; the modeling is:

1. *Comprehensive* in the sense that it aims straight for an overall "inflation rate" encompassing all items.
2. *Atomistic*: every item is modelled individually, having its "private" parameter, its own rate of inflation [$\exp(\pi_t + \beta_i)$], apart from all other items.
3. *Time isolated*: price relatives modeling for period $t-1$ to t is disjoint from that for period $t-2$ to $t-1$ etc.

It is the combination of these suppositions that yields Diewert's "over-parameterized" argument. The primary thrust of Keynes' criticism is against 1: an overall inflation rate or rise/fall in the cost of living has to be a weighted mixture of several price trends. This may be granted without going so far as to embrace item 2. Item 2 is tacitly accepted in almost all (non-stochastic) constructions of price indices. However, it is not at all clear that every single item has its unique price trend. Different items (for example, Brand X ice cream at several supermarkets) are likely to have a tendency to rise and fall together (at least in the long run). There are degrees of homogeneity between

items. In any case, none of these assumptions is a necessary component of a stochastic view of price indices.

4.2 An Elementary State Space Model

We divide the time period t' to t into sub-periods t' , $t' + 1, \dots, t - 1, t$, and the collection of heterogeneous items into homogeneous sub-groups g , where the defining characteristic of homogeneity is a tendency to similarity of price change behavior. We make two assumptions:

1. $I_{t'}$ is a mixture of "homogeneous" indices $I_{gt'}$;
2. $I_{gt'}$ can be attained through chaining: $I_{gt'} = \prod_{\tau=t'+1, \dots, t} I_{g\tau-1, \tau}$, where $\tau = t' + 1, \dots, t$.

We focus on a single group index $I_{gt'}$, dropping the subscript g for simplicity of notation. Thus, for the remainder of this paper, we focus on the "sub-index" $I_{t'}$.

We proceed to develop an elementary state space model (Harvey 1990, Chapter 3) for the logarithms of the within-group price relatives. Suppose the group contains n items. For $i = 1, \dots, n$, let $r_{it} \equiv p_{it}/p_{t-1, i}$ be the micro-period price relatives, and $y_{it} \equiv \log(p_{it}/p_{t-1, i}) = \log(p_{it}) - \log(p_{t-1, i})$, their logs. The reason for using logs is that considerable empirical work, beginning with Edgeworth (see Diewert (1995)), suggests that the logs of price relatives will be much closer to having a normal distribution than the price relatives themselves, which can be considerably skewed. Normal distribution of errors is a standard assumption in state space models. Let $y_t \equiv (y_{t1}, \dots, y_{tn})$ and $\mathbf{1}$ be a vector of ones of length n .

Consider the multivariate random walk plus noise (RWPN) model

$$y_t = \mathbf{1}\mu_t + \varepsilon_t, \quad \varepsilon_t \sim \text{MVN}(0, \sum_{\varepsilon\varepsilon})$$

$$\mu_t = \mu_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_{\eta\eta}) \quad (3)$$

with $\varepsilon_t, \eta_t, \tau \in (t', t' + 1, \dots, t - 1, t)$ mutually independent. The model implies that the amount that overall group prices are rising (or falling) in one micro-period, tends to hover around the amount they tended to rise (or fall) in the previous micro-period. This is a matter of common observation: if the price rise in one month tends to be high (low), then in the next month it tends to be correspondingly high (low). Since we are considering a homogeneous set of items, it makes sense that their log price relatives have a common mean. We leave for later work, the question of how to join sub-indices into an overall index.

The model (3) implies the simpler univariate RWPN model

$$\bar{y}_t = \mu_t + \bar{\varepsilon}_t, \quad \bar{\varepsilon}_t \sim N(0, \sigma_{\bar{\varepsilon}\bar{\varepsilon}})$$

$$\mu_t = \mu_{t-1} + \eta_t, \quad \eta_t \sim N(0, \sigma_{\eta\eta}) \quad (4)$$

with $\bar{y}_t = n^{-1} \mathbf{1}' y_t$, $\bar{\varepsilon}_t = n^{-1} \mathbf{1}' \varepsilon_t$, and $\sigma_{\bar{\varepsilon}\bar{\varepsilon}} = n^{-1} \mathbf{1}' \sum_{\varepsilon\varepsilon} \mathbf{1}$. Some information is thrown away in using (4); on the other hand, the normality assumption is even more likely to hold. For convenience, calculations in the study described in Section 5, were based on the univariate model.

The Kalman Filter (Harvey 1990, Section 3.2) can be used to give estimates $\hat{\mu}_t$, and $\hat{\sigma}_{\bar{\varepsilon}\bar{\varepsilon}}, \hat{\sigma}_{\eta\eta}$ of the state parameters μ_t and the variances $\sigma_{\bar{\varepsilon}\bar{\varepsilon}}, \sigma_{\eta\eta}$ respectively.

Then we define $I_{t'} \equiv E(G_{t'} | S_t)$, where $G_{t'} = \prod_i (p_{it}/p_{t-1, i})^{f_i}$ is a geomean dependent on fixed shares f_i , and S_t represents the totality of state parameters μ_t through time t , and also the "hyperparameters" $\sigma_{\bar{\varepsilon}\bar{\varepsilon}}, \sigma_{\eta\eta}$. In other words, we condition on what we take to be the underlying process through time t . Then

$$I_{t'} = \exp\left(\mu_t + \mu_{t-1} + \dots + \mu_{t'+1} + \frac{1}{2} v\right), \quad (5)$$

where $v = (t - t') \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'}$, with $\sigma_{ii'}$ the covariance of ε_{ii} and $\varepsilon_{i'i'}$ typically of lower order than the state parameters μ_{ti} . The natural estimator of $I_{t'}$ is $\hat{I}_{t'} \equiv \exp(\hat{\mu}_t + \hat{\mu}_{t-1} + \dots + \hat{\mu}_{t'+1})$; then

$$E(\hat{I}_{t'} | S_t) = \exp\left(\mu_t + \mu_{t-1} + \dots + \mu_{t'+1} + \frac{1}{2} \tilde{v}\right), \quad (6)$$

where \tilde{v} , given in the Appendix, does not in general equal v , but is frequently close, and in any case is of the same order of magnitude. The difference $\Delta(v) = \tilde{v} - v$ can be estimated, by say $\hat{\Delta}(v)$, yielding a bias-corrected estimator $\tilde{I}_{t'} \equiv \hat{I}_{t'} \exp(-1/2 \hat{\Delta}(v))$. Expressions for v and \tilde{v} , and a suggestion for a maximal $\hat{\Delta}(v)$, are given in the Appendix. It may be noted that $\hat{\Delta}(v)$, and hence $\tilde{I}_{t'}$, depends on the weights f_i , but that $\hat{I}_{t'}$ does not.

5. EMPIRICAL STUDY

To determine the feasibility of the calculation of price indices using the RWPN model and gain some idea of the behavior of the RWPN index, a small empirical study was made, using price and quantity data for Canned Tuna in the A.C. Nielsen Academic Data Base. Canned tuna has somewhat volatile price and quantity behavior, due to frequent sales, at sometimes very marked discount.

The study covered the Northeast USA and the 104 weeks of the years 1992-1993. The original data set was rather large. To make the investigation manageable, weekly data was combined into 4-week periods, giving a total of 26 periods over two years. Thus for purposes of this study, the data were cumulative quantities and quantity-weighted average prices over four week periods.

The homogeneous groups were defined by brand and type, as follows: 3 brands here labeled A, B, C of "premium" tuna in water, the same three brands of "light" tuna in oil, and again the same three brands of "light" in water, making 9 groups in all.

The study focused on 83 outlets which had positive quantities over most of the 4 week periods, for each of the 9 distinct groups.

The RWPB based index $\hat{I}_{t,t'}$ and the adjusted RWPB based index $\tilde{I}_{t,t'}$ were calculated for four time intervals. In each case, the final period $t = 26$, and the early period was taken successively as $t' = 3, 6, 10, 14$. For the purpose of comparison, we also calculated the corresponding Laspeyres and Paasche Indices. These two standard indices provide also a basis of indirect comparison to the Fisher and Törnqvist, which will be about half way between them.

Figures 1 and 2, for premium and light tuna respectively, give the values of the four indices for the four time intervals, the points representing the state space indices, the lines used to indicate the Laspeyres and Paasche. The adjusted RWPB $\tilde{I}_{t,t'}$ is invariably larger than the unadjusted RWPB $\hat{I}_{t,t'}$. Note that, since it is the first period that we are varying, where the path of indices is monotone up, this would suggest a downward trend in the cost of the particular tuna group (and vice versa).

We observe that the new indices are not out of line with the traditional indices, frequently lying between the Laspeyres and Paasche, but they tend to be considerably more stable as t' changes, suggesting possibly that the traditional indices are reacting to "noise" in the data, and that, in fact, basically very little change is going on in this

two year period. It may also be observed in Figure 2, that Light in Oil and Light in Water have similar within brand behavior, suggesting that we might have taken a broader "homogeneous" grouping.

6. FURTHER WORK

The investigation described in this paper suggests several topics for further research.

Measures of precision and estimates of the RWPB indices, in terms of variances or confidence intervals based on the state space model, need to be worked out. Even those who are dubious about the viability of a stochastic methodology in price indices, find the possibility of having a measure of precision appealing (Diewert 1995). It would also be of interest to get measures of precision of more standard indices, based on the state space model.

Empirical work is desirable that investigates more closely what groups of items might best be considered "homogeneous". Also, models possibly more elaborate than the simple RWPB model require investigation. In this respect, the use of scanner data will be a great help, supplying as it does, quantity data as well as prices, in great detail.

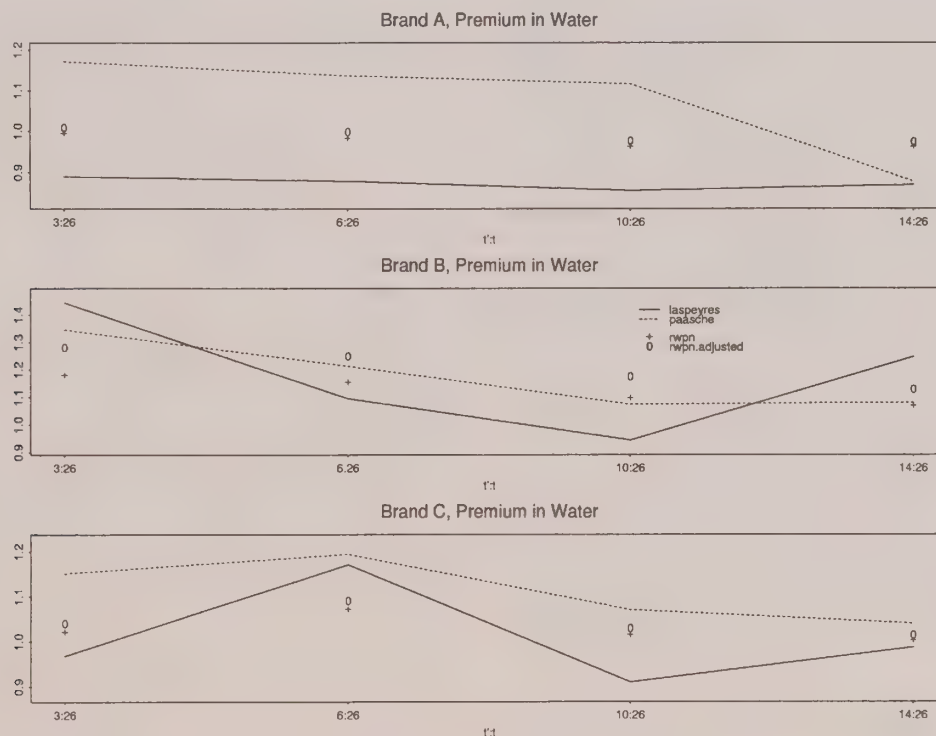


Figure 1. Four Price Indexes for Four Time Intervals, Premium Tuna

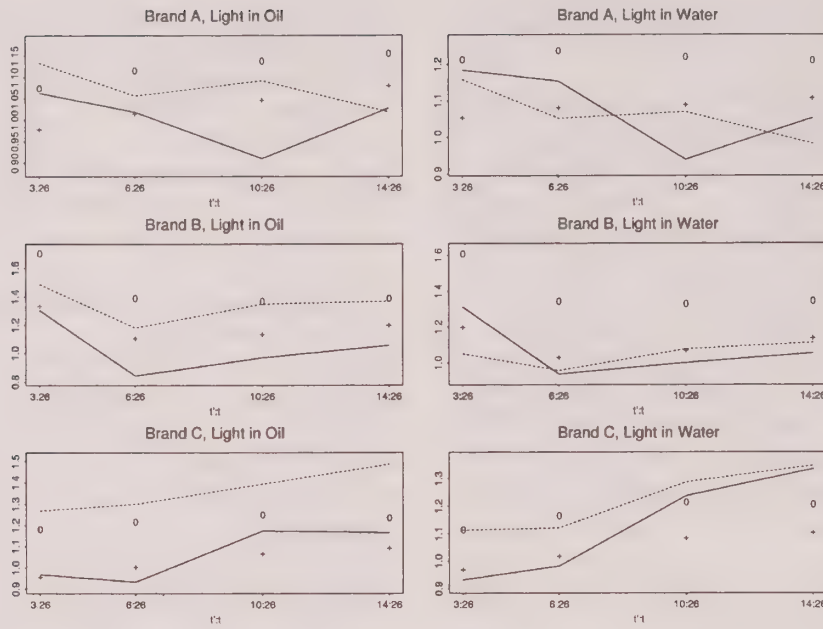


Figure 2. Four Price Indexes for Four Time Intervals, Light Tuna

The state space methodology has methods of handling missing data (Harvey 1990, Section 3.4.7). A point of major concern is how well these models will handle missing data in estimating price indices. In particular, since in practice most data for calculating price indices is based on a small sample of items available, we need to know the robustness of state space indices to the absence of data.

Algorithms for smoothing and forecasting of state space models, are well known. Their use in revising and forecasting indices, might be of great interest.

Finally, in this paper we have focussed only on getting an index for a single homogeneous group. It would be of interest to develop a state space model that combines groups and enables us to get an overall measure of purchasing power.

ACKNOWLEDGEMENTS

The author thanks B. Moulton, S. Scott, M. Reinsdorf, R. Tiller, B. Balk, and J. Aldrich for discussions of the ideas in this paper.

APPENDIX

Details of expressions (5) and (6).

We have that

$$G_{t't} = \prod_i \left(\frac{p_{ti}}{p_{t-1,i}} \frac{p_{t-1,i}}{p_{t-2,i}} \dots \frac{p_{t'+1,i}}{p_{t'i}} \right)^{f_i}$$

$$= \prod_i (r_{ti} r_{t-1,i} \dots r_{t'+1,i})^{f_i},$$

and letting

$$H_{t't} = \log(G_{t't}) = \sum_i f_i \log(p_{ti}/p_{t'i}),$$

we have that

$$H_{t't} = \sum_i f_i \log(r_{ti} r_{t-1,i} \dots r_{t'+1,i})$$

$$= \sum_i f_i (y_{ti} + y_{t-1,i} + \dots y_{t'+1,i})$$

and also that

$$I_{t't} = E(G_{t't}) = \exp(E(H_{t't}) + 1/2 \text{ var}(H_{t't})),$$

where the moments are calculated conditional on the state S_t , as in Section 4.3. Let $v = \text{var}(H_{t'})$. Then

$$E(H_{t'}) \equiv E(H_{t'} | S_t) = \sum_i f_i (\mu_t + \mu_{t-1} + \dots + \mu_{t'+1}) = \mu_t + \mu_{t-1} + \dots + \mu_{t'+1}$$

and

$$v = \text{var}(H_{t'}) \equiv \text{var}(H_{t'} | S_t) = \text{var} \left(\sum_{\tau=t'+1}^t \sum_i f_i \varepsilon_{ti} | S_t \right) = (t - t') \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'},$$

where $\sigma_{ii'}$ is the covariance of ε_{ti} and $\varepsilon_{ti'}$. We note that $v = (t - t') \sum_i f_i^2 \sigma_{ee}$, in the special case that the errors ε_{ti} are independent and identically distributed at each time period.

We now consider estimator $\hat{I}_{t'} \equiv \exp(\hat{\mu}_t + \hat{\mu}_{t-1} + \dots + \hat{\mu}_{t'+1})$. We find that $E(\hat{I}_{t'}) = \exp(\mu_t + \mu_{t-1} + \dots + \mu_{t'+1} + 1/2 \tilde{v})$, where

$$\tilde{v} \equiv \text{var} \left(\sum_{\tau=t'+1}^t \hat{\mu}_\tau | S_t \right) = \left\{ \sum_{\tau=t'+1}^t \gamma_\tau^2 \right\} \text{var}(\bar{y}_t | S_t) + \gamma_{t'}^{*2} \text{var}(\hat{\mu}_{t'} | S_t) = \left\{ \sum_{\tau=t'+1}^t \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \right\} \sigma_{ee}$$

with

$$\gamma_\tau = k_\tau \left(1 + \sum_{v=\tau+1}^t \prod_{u=\tau+1}^v (1 - k_u) \right)$$

and

$$\gamma_{t'}^* = \sum_{v=t'+1}^t \prod_{u=t'+1}^v (1 - k_u),$$

where

$$k_\tau = p_{\tau|\tau-1} / (p_{\tau|\tau-1} + 1),$$

and $p_{\tau|\tau-1}$, p_τ are the mean square errors of $\hat{\mu}_\tau$ given data up to $\tau - 1$, τ respectively, and are estimated using the Kalman Filter.

This result follows from the equations used in estimating μ_τ :

$$\begin{aligned} \hat{\mu}_t &= k_t \bar{y}_t + (1 - k_t) \hat{\mu}_{t-1} \\ \hat{\mu}_{t-1} &= k_{t-1} \bar{y}_{t-1} + (1 - k_{t-1}) \hat{\mu}_{t-2} \\ &\vdots \\ \hat{\mu}_{t'+1} &= k_{t'+1} \bar{y}_{t'+1} + (1 - k_{t'+1}) \hat{\mu}_{t'} \end{aligned}$$

(cf. Harvey 1990, equation 3.2.8), by expressing each $\hat{\mu}_\tau$ in terms of $\bar{y}_\tau, \bar{y}_{\tau-1}, \dots, \bar{y}_{t'+1}, \hat{\mu}_{t'}$.

In comparing v and \tilde{v} , we find, empirically that

$$\sum_{\tau=t'+1}^t \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \approx t - t'.$$

We here consider the simple case where $\text{var}(\varepsilon_{ti}) = \sigma_{ee}$ and $\text{cov}(\varepsilon_{ti}, \varepsilon_{ti'}) = \rho \sigma_{ee}$, with $\rho \geq 0$, for $i' \neq i$, that is where not only variances, but all covariances are equal and non-negative. It then can be shown that

$$\sigma_{ee} = n^{-2} \sum_i \sum_{i'} \sigma_{ii'} \leq \sum_i \sum_{i'} \sigma_{ii'} f_i f_{i'} \leq n \sum_i f_i^2 \sigma_{ee},$$

where n is the number of items in the group. The lower bound is achieved in the case $f_i = 1/n$, and the upper in the case $\rho = 0$. In the first case, no bias adjustment is necessary; in the second, we would take $\hat{\Delta}(v) = \hat{v} - \hat{v}$, where $\hat{v} = (t - t') n \sum_i f_i^2 \hat{\sigma}_{ee}$ and $\hat{v} = \{ \sum_{\tau=t'+1}^t \gamma_\tau^2 + \gamma_{t'}^{*2} p_{t'-1} \} \hat{\sigma}_{ee}$. These correspond respectively to $\hat{I}_{t'}$ and $\hat{I}_{t'}$.

REFERENCES

- ALDRICH, J. (1992). Probability and depreciation: a history of the stochastic approach to index numbers. *History of Political Economy*, 24, 657-87.
- BALK, B.M. (1980). A method for constructing price indexes for seasonal commodities. *Journal of the Royal Statistical Society, A*, 143, 68-75.
- BALK, B.M. (1995). Axiomatic price theory: a survey. *International Statistical Review*, 63, 69-93.
- BRYAN, M.F., and CECCHETTI, S.G. (1993). The consumer price index as a measure of inflation. *Economic Review, Federal Reserve Bank of Cleveland*, 29, 15-24.
- CLEMENTS, K.W., and IZAN, H.Y. (1981). A note on estimating Divisia index numbers. *International Economical Review*, 22, 745-747.
- CLEMENTS, K.W., and IZAN, H.Y. (1987). The measurement of inflation: a stochastic approach. *Journal of Business and Economic Statistics*, 5, 339-350.
- DIIEWERT, W.E. (1987). Index numbers. In *The New Palgrave: A Dictionary of Economics*, (Eds. J. Eatwell, M. Milgate, and P. Newman). London: MacMillan.
- DIIEWERT, W. E. (1995). On the Stochastic Approach to Index Numbers. Discussion Paper No. DP 95-31, Department of Economics, University of British Columbia.
- EICHHORN, W., and VOELLER, J. (1976). *Theory of the Price Index*. Berlin: Springer-Verlag.
- FISHER, I. (1922). *The Making of Index Numbers*. Boston: Houghton Mifflin.
- HARVEY, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
- KEYNES, J.M. (1930). *A Treatise on Money*. New York: Harcourt, Brace and Company.
- KOTT, P.S. (1984). A superpopulation approach to the design of price index estimators with small sampling biases. *Journal of Business and Economic Statistics*, 2, 83-90.
- SELVANATHAN, E.A., and RAO, D.S.P. (1994). *Index Numbers: A Stochastic Approach*. Ann Arbor: The University of Michigan Press.

Treatment of Nonresponse in Cycle Two of the National Population Health Survey

JEAN-LOUIS TAMBAY, IOANA ȘCHIOPU-KRATINA, JACQUELINE MAYDA,
DIANA STUKEL and SYLVAIN NADON¹

ABSTRACT

The National Population Health Survey (NPHS) is one of Statistics Canada's three major longitudinal household surveys providing an extensive coverage of the Canadian population. A panel of approximately 17,000 people are being followed up every two years for up to twenty years. The survey data are used for longitudinal analyses, although an important objective is the production of cross-sectional estimates. Each cycle panel respondents provide detailed health information (H) while, to augment the cross-sectional sample, general socio-demographic and health information (G) are collected from all members of their households. This particular collection strategy presents several observable response patterns for Panel Members after two cycles: GH-GH, GH-G*, GH-**, G*-GH, G*-G* and G*-**, where "**" denotes a missing portion of data. The article presents the methodology developed to deal with these types of longitudinal nonresponse as well as with nonresponse from a cross-sectional perspective. The use of weight adjustments for nonresponse and the creation of adjustment cells for weighting using a CHAID algorithm are discussed.

KEY WORDS: Longitudinal surveys; Treatment of nonresponse; CHAID algorithms.

1. INTRODUCTION

In 1996-97, Statistics Canada completed data collection for Cycle 2 of the National Population Health Survey (NPHS). This longitudinal survey was launched in 1994 to provide comprehensive information on the health status of the Canadian population and on the determinants of health. The in-scope population covers residents of households and health institutions throughout the country. In the provinces the household questionnaire has two main components which are administered using computer-assisted interviewing. The General component collects socio-demographic and basic health information for each member of the household. The Health component obtains more detailed health information about the household member selected to participate in the longitudinal panel.

Although the NPHS is a longitudinal survey, its objectives also include the production of periodic cross-sectional estimates (Catlin and Will 1992). The data collection methodology reflects both longitudinal and cross-sectional needs. Panel Members, chosen in Cycle 1, are followed-up every two years for up to 20 years. Persons residing with the Panel Members at those times provide General component information for use in cross-sectional estimation. As the cross-sectional coverage of the sample deteriorates over time, the sample needs to be "topped-up" periodically. The first top-up is planned for Cycle 3, in 1998.

This paper presents the methodology developed in Cycle 2 to deal with nonresponse at the household and person levels (flagging will be used for item nonresponse). The methodology is based on reweighting respondents within

sub-populations called weighting cells to account for nonresponse. Reweighting is a common approach for the treatment of item nonresponse. The bias and variance of this approach have been considered by Thomsen (1973), Oh and Scheuren (1983), Kalton and Kasprzyk (1986) and Little (1986), among others. If weighting cells are defined such that nonresponse occurs almost completely at random within each cell then the bias due to nonresponse can become negligible. In a similar vein David, Little, Samuhel and Triest (1983) extended to nonresponse the theory developed by Rosenbaum and Rubin (1983) in the context of propensity score matching in observational studies. Their results imply that reweighting can adjust for nonresponse bias when the weighting cells are formed based on the propensity to respond.

An overview of the NPHS sample design and outputs for the first two Cycles is given in Section 2. Section 3 presents the nonresponse treatment strategies and their results. Concluding remarks are given in Section 4. Note that the methodology presented pertains to the provincial household samples; it does not cover the samples in the territories and in institutions.

2. OVERVIEW OF THE NPHS DESIGN AND OUTPUTS

2.1 Cycle 1 Sample Design

The initial sample of households was selected in 1994 using the sample selection vehicle built for the Canadian Labour Force Survey (LFS), and, in the province of Quebec, using dwellings that had participated in a health survey

¹ Jean-Louis Tambay, Ioana Șchiopu-Kratina, Jacqueline Mayda, Diana Stukel and Sylvain Nadon, Household Survey Methods Division, Statistics Canada, 16th floor, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

conducted by Santé Québec the previous year. In both cases the households or dwellings were selected at random within stratified samples of clusters selected using probability proportional to size. The clusters were organized into replicates and collection period to capture seasonality and for variance estimation purposes. There were two "summer" collection periods (June and August) and two "winter" collection periods (November and March, 1995).

Figure 1 illustrates the Panel selection mechanism applied outside the province of Quebec. Sample households were randomly designated as "Adult" or "Children" households, and as eligible for screening or not, prior to collection. Screening increased the presence in the panel of inhabitants of larger households who would be under-represented with the selection of only one member per household, particularly children and youths. Households eligible for screening were rejected from the sample if they had no member aged under 25. Screening was not used in Quebec as information from the provincial health survey allowed the application of different sub-sampling rates by household type and size.

Sample Unit Type	Household Characteristic	Panel selection restricted to:
"Children" household Eligible for Screening (EFS)	No member under 25	N/A – hhld rejected
	No children, some members under 25	Any member
	Children present	Child members
"Children" household not EFS	No children present	Any member
	Children present	Child members
"Adult" hhld	All	Members over 12

Figure 1. Panel Selection Mechanism Outside Quebec

The classification into "Adult" and "Children" households was done for an operational reason: the Health questionnaire for children, would not be available before the winter collection periods. In "Adult" households, which could be interviewed any time, children under 12 were not eligible for the panel. "Children" households, even those in "summer" clusters, were interviewed in a winter collection period. If children were present in those households then the panel selection was restricted to them. To diminish the seasonal distortions to the data collection workload and the panel representability brought about by these procedures, fewer households were classified as "Children" households in summer clusters, and, with one minor exception, screening was applied only to "Children" households.

Provinces wishing to improve sub-provincial estimates could fund additional sample sizes. In three provinces this was done by augmenting the sample size in targeted regions. In British Columbia an additional sample of about

800 households was selected in a local health unit using Random Digit Dialling (RDD). The expected total sample size in the provinces was approximately 23,000 households after screening.

The above gives a general indication of the 1994 sample design which is sufficient for the needs of this paper. Readers wishing a more precise presentation of the 1994 sample should see Tambay and Catlin (1995), or Statistics Canada (1995).

2.2 Cycle 1 Weighting and Outputs

The major output of the NPHS consists of person-level anonymized Public Use Microdata Files (PUMFs) of survey responses (internal versions of those files that include information suppressed for confidentiality reasons are also created). For 1994 a General PUMF (58,400 records) and a Health PUMF (17,600 records) were released containing the General and Health information collected from every household member and from the selected non-child Panel Members, respectively (Statistics Canada 1995).

The sample weights attached to every record on the PUMF were calculated by applying a series of adjustments to a basic weight representing the household inverse sampling rates (ISR). The ISRs are calculated by multiplying the weights of the original LFS or Santé Québec samples by the inverse of the sub-sampling rates applied by the NPHS. For the sake of brevity we will only describe the main adjustments used outside of Quebec.

Adjustments to the weights for the General PUMF include: (1) a household nonresponse adjustment; (2) an adjustment for the rejective method; (3) an adjustment for person nonresponse [within responding households] and, finally; (4) a simple post-stratification adjustment. Adjustment (2) was applied only to households with no member under 25. It was $1/(1 - r_s)$, where r_s was the sub-sampling rate for the screening applied in the stratum. The post-stratification adjustment was done separately for each province-age group-sex cross-class. Weights resulting from all earlier steps are multiplied by the ratio of known to estimated population sizes within the cross-class. The known population sizes are in fact Census-based projections.

The adjustments for household and person level nonresponse (at 11.3% and 1.4%, respectively) were applied to respondent units as the nonrespondents were excluded from the PUMFs. If w_i is the sample weight of a unit i , the nonresponse-adjusted weight, $w_{adj,i}$ is defined as $w_{adj,i} = w_i (\sum_{all} w_i) / (\sum_{resp} w_i)$, where the sums are taken over all sample units and all respondent units, respectively, within nonresponse adjustment weighting cells. Due to a lack of information on nonrespondent households the weighting cells for household level nonresponse were simply cross-classes of NPHS strata and season (*i.e.*, "summer" vs. "winter" clusters). Weighting cells for the person level nonresponse, which was very low, were the province-age-sex cross-classes that were used for the post-stratification adjustment.

Adjustments to the weights for the Health PUMF included: (1) a household nonresponse adjustment; (2) an adjustment for the rejective method; (3) an adjustment for the "Adult/Children" household sub-sampling; (4) an adjustment for the longitudinal Panel Member selection; (5) an adjustment for Panel Member nonresponse; and (6) a post-stratification adjustment. The first two adjustments were exactly those for the General PUMF. As the Health PUMF did not include Panel Members who were children, adjustment (3) compensated for those sample households where non-children were ineligible for panel membership. The adjustment thus applied only to households with children and was equal to $1/r$, where r was the proportion of "Adult" households in the sample. Adjustment (4) was the inverse of the probability of having selected the Panel Member. The adjustments for Panel Member nonresponse (at 3.9%) and for post-stratification were similar to those for the General PUMF, and used the same province-age-sex cross-classes. Although child Panel Members were not included in the Health PUMF, for longitudinal purposes their sample weights were obtained as above using $1/(1-r)$ instead of $1/r$ in step (3).

2.3 Cycle 2 Sample Design

In Cycle 2 the focus of the survey was more on longitudinal estimation: no sample "top-up" was planned until the following cycle. The "Core" sample thus consisted of about 17,000 Panel Members and their current households. Panel Members were traced and administered the General and Health questionnaire components, while other members of their household were administered the General component only. No follow-up was done for 1994 nonresponding households. In Alberta, Manitoba and Ontario large (non-Core) additional samples were obtained, using RDD, to allow the production of cross-sectional estimates at sub-provincial levels. In every RDD household one member aged over 12 was selected to complete the Health component. In Alberta and Manitoba, RDD households with children also had a child selected to complete the Health component.

We note that, for cross-sectional purposes, the Core sample does not cover very well arrivals in the population such as newborns and recent immigrants. The population administered the General questionnaire consists of residents of households where at least one member was in-scope in Cycle 1; households made up entirely of recent immigrants (and their newborns) are thus missed. The population administered the Health questionnaire consists of persons who were in-scope in Cycle 1: recent immigrants and children under 2 years old are excluded from the Core target population (they are included in the RDD target population). For both the General and the Health questionnaires post-stratification is done using population figures that do not exclude the recent immigrants. The result is that the population of recent immigrants is implicitly being estimated for by the population of non-

immigrants because the latter's Core weights are adjusted upwards to account for the former's numbers. This is a limitation that is acknowledged in the PUMF documentation. Alternative methods would have been to post-stratify using only non-immigrant population projections or to somehow adjust only the weights of less recent immigrants (who are covered) to account for the more recent immigrants (who are not). These methods would have been difficult to apply where, for the General questionnaire, a distinction between immigrants in immigrant-only households and immigrants in mixed households would have been required.

2.4 Cycle 2 Weighting and Outputs

Figure 2 summarizes the survey's three major outputs planned for Cycle 2: a Longitudinal PUMF; a Health Cross-Sectional PUMF and a General Cross-Sectional PUMF. The planned Longitudinal PUMF contains General and Health information for both Cycles for the 17,000 Panel respondents [note: confidentiality requirements may prevent the release of a longitudinal PUMF – in which case only an internal microdata file will be produced]. The Health Cross-Sectional PUMF contains 1996 General and Health information for about 70,000 Panel Members and RDD Selected Members. The General Cross-Sectional PUMF contains 1996 General information for about 220,000 members of the Core and RDD samples. The weighting processes involved for each PUMF, presented below for the Core sample, are described in more detail in Stukel, Mohl and Tambay (1997).

Output File	LONGITUDINAL PUMF	HEALTH CROSS-SECTIONAL PUMF	GENERAL CROSS-SECTIONAL PUMF
Contents	General & Health	General & Health	General only
Samples	Core only	Core & RDD (3 provs.)	Core & RDD (3 provs.)
Units	Panel Member (PM)	PM/RDD Sel. Mem.	All Hhld. Members
Size	≈ 17,000 records	≈ 70,000 records	≈ 220,000 records
Weighting Strategy (for Core Sample)	1.Base Year Weight 2.PM Nonresp. Adjustment 3.Post-stratification	1.Base Year Weight 2.PM Nonresp. Adj. 3.Core/RDD integration 4.Post-stratification	1.Base Year Weight 2.Hhld. Nonresp. Adj. 3.Weight Share Adj. 4.Hhld. Mem. NR Adj. 5.Core/RDD integration 6.Post-stratification

Figure 2. Description of Output Files for Cycle 2

Respondent survey weights on the Longitudinal PUMF are obtained by adjusting a base year weight first for 1996 panel nonresponse and then for post-stratification. The base year weight represents the inverse sampling rate for 1994 including all Health PUMF adjustments described in section 2.2 up to adjustment (4) for panel selection (a correction is needed for the "removal" of the 1994 provincial sample additions). The weight adjustment for

nonresponse is the focus of the following section and will be described there. Post-stratification is done to reproduce 1994 provincial population counts by age-sex categories.

For the Health Cross-Sectional PUMF, the weighting process for (Core) Panel Members involves three or four steps. Usually, the base year weight is adjusted for Panel Member nonresponse, as explained in the following section, and for post-stratification (to match 1996 provincial or regional population counts by age-sex categories). In provinces with RDD samples the extra step is the integration with the RDD sample. The integrated estimate is obtained by a somewhat degenerate adaptation of the Skinner-Rao dual frame estimator (Skinner and Rao 1996).

For the General Cross-Sectional PUMF, the weighting process for the core sample involves five or six steps. First, once more, is the calculation of the base year weight. Then comes an adjustment for nonresponse at the household level, discussed in the following section. The next step is the application of the “weight share method”. The method was described by Ernst (1989) and developed further by Lavallée (1995). The Panel Member’s weight, divided by the number of persons in his/her household who were in-scope in Cycle 1, is assigned to all household members including those who were not in-scope in Cycle 1 (*e.g.*, births, immigrants). The method is unbiased for estimates of totals for the population of households where at least one member was in-scope in Cycle 1. The next step is a household member nonresponse adjustment. In RDD provinces this is followed by integration of the Core sample with the RDD sample (this time for all ages). Post-stratification is done in a similar fashion to that for the Health Cross-Sectional PUMF.

3. CYCLE 2 CORE SAMPLE NONRESPONSE STRATEGY

This section presents the strategy adopted for the treatment of Cycle 2 nonresponse in the Core (non-RDD) sample. Adjusting for nonresponse was done once again using the weighting cell approach except that, this time, Cycle 1 data were available to create weighting cells that are more homogeneous with respect to the propensity to respond, and thus more apt to remove nonresponse bias. Section 3.1 identifies nonrespondents in the NPHS. Section 3.2 discusses two general approaches for the creation of weighting cells, giving the one chosen for the NPHS. The strategy for the adjustment for nonresponse is explained in section 3.3 while section 3.4 describes the creation of the nonresponse weighting cells.

3.1 Definitions of Nonrespondent and Out-of-scope Units

The first step in the treatment of nonresponse consisted of its definition or identification. In Cycle 2, questionnaires were fully completed for 89% of the Core sample and

partially completed for another 3%. The rest of the sample consisted of refusals (3.1%), of cases where the Panel Member could not be traced (1.7%), had died (1.7%), had left Canada (0.5%), or was institutionalized (0.4%), and of other types of nonresponse such as temporary absences and special circumstances (0.7%). Within responding households person level nonresponse was very low: 1.8% for the General questionnaire and 1.1% for the Health questionnaire. We first identify cases that are not nonresponses for longitudinal and cross-sectional purposes.

For longitudinal purposes a death is considered a valid survey response. Panel Members who had died before Cycle 2 had their status recorded as such and the 1996 portion of their data coded as “Not Applicable” on the Longitudinal microdata file. Panel Members who moved to an institution or to the Territories were followed-up and their responses were used for longitudinal purposes. Panel Members who left the country were not followed-up but were treated as longitudinal nonrespondents even though it would have been more accurate for some analyses to have considered them as having left the scope of the study. This treatment was chosen because such persons would fall back in-scope should they move back to Canada.

For cross-sectional purposes all the cases presented in the preceding paragraph were treated as out-of-scope situations. This was acceptable because the separate Institutional and Territorial survey vehicles assumed the cross-sectional coverage of these particular in-scope populations. Out-of-scope units were not on the PUMFs but, as they represented other such units, they were treated for weighting purposes like respondents in all the weight adjustment steps except the integration and post-stratification steps.

Refusals and cases where questionnaires were missing for reasons other than those given in the preceding paragraphs were defined as nonresponses. As will be seen, a distinction was later made between “full” and “partial” longitudinal nonrespondents to accommodate different users.

3.2 Approach for Creating Nonresponse Adjustment Weighting Cells

Two statistical approaches for creating response weighting cells involve segmentation modelling and logistic regression. An example of the latter is given in Czajka, Hirabayashi, Little and Rubin (1992). The authors obtained advance taxation estimates from early tax filer returns using adjustment weighting cells that were based on ranges of propensities to file early. Logistic regression was used to estimate tax filers’ propensities to file early. The longitudinal Survey of Labour and Income Dynamics (SLID) provides another example involving logistic regression (Grondin 1996). Sample units’ response indicators were regressed on known (dichotomous) characteristics. Adjustment cells for nonresponse were generated by cross-classifying the sample units using all the significant covariates. In order to respect minimum cell sizes and response rates some collapsing was done starting with cells sharing all but the least significant covariates.

In the segmentation modelling approach a decision tree structure is generated from the data by successively splitting the data set such that, at each node, the most significant predictor for the response variable is used to define the following split. The splitting continues until one cannot find any significant variable for the split or minimum cell size requirements cannot be respected. An early application of segmentation modelling for nonresponse adjustment was with respect to the Panel Study of Income Dynamics (Institute for Social Research 1979). Because of its advantages, given below, the NPHS adopted the segmentation modelling approach using the CHAID algorithm developed by Kass (1980). The CHAID (Chi-square Automatic Interaction Detection) algorithm uses χ^2 tests to define splits for categorical predictors and retains the most significant split at each stage. The splitting, into two or more categories, is done differently for ordered and unordered predictors. CHAID was applied using the Knowledge Seeker software program (ANGOSS Software 1995). Note that Knowledge Seeker applies CHAID to continuous predictors by first transforming them into ordered discrete variables.

Advantages and disadvantages of the logistic and CHAID approaches are known and documented (for example see Kalton and Kasprzyk 1986). The logistic regression approach is based on theory familiar to many analysts, and can be programmed using a number of standard statistical packages. It also provides individual estimates of response propensity that can be used directly to adjust the weights of respondents. However, to ensure reasonable program execution times the number of variable and interaction terms used must usually be limited. Collapsing cells can also become complicated, as in the case of SLID above. The CHAID algorithm offers the advantages of accepting a large number of covariates and, by its decision tree structure, easily accommodating interactions among them. Moreover, minimum cell size requirements can easily be incorporated as program execution parameters. Its main disadvantages are a less familiar theoretical underpinning (Knowledge Seeker is advertised as a "data mining" tool) and the limited documentation and software available for its implementation. It should also be mentioned that, while some statistical packages such as SUDAAN and PC CARP can incorporate the sample design when fitting logistic models to survey data, this is not the case with CHAID. The NPHS tried to address this limitation by including as predictor variables characteristics that were related to the sample design (see Section 3.4).

Two empirical studies comparing the logistic and CHAID approaches for the treatment of nonresponse obtained different results. Rizzo, Kalton and Brick (1996) did not find much of a difference between the two approaches for the Survey of Income and Program Participation. On the other hand Dufour, Gagnon, Morin, Renaud and Särndal (1998), in a simulation study for SLID, obtained a lower bias after nonresponse adjustment with the CHAID approach.

3.3 Adjusting for Nonresponse in the Core Sample

Nonresponse adjustments had to be developed for each PUMF: Longitudinal, General (Cross-Sectional) and Health (Cross-Sectional). We will deal with the General PUMF first.

As Figure 2 showed, the weighting strategy for the General PUMF required separate adjustments for nonresponse at the household and at the person levels. In creating adjustment cells for household level nonresponse, characteristics of the Panel Member as well as those of the household were considered as nonresponse predictors. This was done for three reasons. Firstly, as the Panel Member was the link to the household in Cycle 2, his or her characteristics may be related to finding the household and obtaining a response (the first contact will often be through him or her). Secondly, a few personal characteristics of the Panel Member, such as race, are in some sense household characteristics. Finally, using Panel Member characteristics was not incompatible with our need to use a variety of information for the construction of weighting cells. If Panel Member characteristics are not significant, then CHAID simply does not retain them.

Person level nonresponse to the General component occurred when the information was available for some but not all of the household members, perhaps due to members' refusals or temporary absences. Given the low 1.8% nonresponse rate at the person level, it was felt that the creation of weighting cells based on province-age-sex categories (as in Cycle 1) would be sufficient for our needs.

In contrast to the General PUMF, the adjustments for household and person level nonresponse for both the Longitudinal and the Health PUMFs could be combined into a single adjustment as they concerned only one person – the Panel Member. A single set of adjustment cells thus needed to be created.

For the Longitudinal PUMF it was noted that the data items came from both the General and Health components but that response rates for the two components were different. This difference produced data with different Cycle 1-Cycle 2 reporting patterns: GH-GH, GH-G*, G*-GH, G*-G*, not to mention longitudinal nonresponse patterns GH-** and G*-, where the letters stand for the component reported each Cycle ("*" if not reported). To maximize the utility of the data it was decided to do two adjustments for longitudinal nonresponse. One adjustment would be for the "Full Longitudinal Response" pattern GH-GH. In other words, all other response patterns would be considered as nonresponses. The other adjustment would be for the "Partial Longitudinal Response" pattern which included cases where, at minimum, General information was available for each cycle (patterns GH-GH, GH-G*, G*-GH and G*-G*). The Full Response data set could be used by researchers who would like to analyse a full longitudinal data set covering the entire questionnaire contents. The Partial Response data set could be of use to researchers primarily interested in the types of variables that are on the

General questionnaire. As the counts in Table 1 below show, the Partial Longitudinal Response data set is only about 3% larger than the Full Longitudinal Response data set.

Table 1 Longitudinal Response Patterns			
Response Type		Cycle 1-2 Response Pattern	Number of records
Full	Partial		
■	■	GH-GH	15,670
	■	GH-G*	110
	■	G*-GH	366
	■	G*-G*	22
		GH-**	1,014
		G*-**	94
Total			17,276

Based upon the above, adjustment cells must be built for five types of responses (or nonresponses) in Cycle 2:

- General PUMF – household response
- General PUMF – person response
- Health PUMF – combined response
- Longitudinal PUMF – full response
- Longitudinal PUMF – partial response

Only three sets of adjustment cells were created for those response types. Adjustment cells created for the General PUMF household level responses were also used for the Longitudinal PUMF partial responses because getting a response from a household led almost always to obtaining a partial response for the longitudinal member (there were 53 exceptions). Likewise, adjustment cells generated for full respondents on the Longitudinal PUMF were used for the Health PUMF responses. Although there were 366 more cases of responses of the latter type (pattern G*-GH) it was considered that the same response mechanism was at work in both instances. The third set of adjustment cells was for person level responses on the General PUMF. Province-age-sex categories were used, as was done in Cycle 1.

Note that, although the same adjustment cells would serve for different data sets, the nonresponse weight adjustments would be calculated separately for each data set type. Thus, the 366 records with response pattern G*-GH would be treated as respondents when adjusting weights for the Health PUMF, but as nonrespondents when adjusting weights for full respondents on the Longitudinal PUMF.

3.4 Creation of Weighting Adjustment Cells

Separate sets of weighting adjustment cells were created for each province. The first step consisted of identifying the variables to consider. With CHAID the number of variables that could be considered was not really an issue, and different types were considered. Characteristics of the household, or dwelling, as well as personal characteristics of the Panel Member would of course be considered. In an effort to incorporate the design of the survey into the

analysis some characteristics that were related to the design of the survey or to the sampling weight were also considered. These included geographical variables such as the Census Metropolitan Area code or the Urban/Rural indicator, special Cycle 1 design variables such as the flag identifying households for screening and the “Adult/Children” household type, and characteristics related to the application of those design variables, such as the presence in the household of a member aged under 25 or of a child. The household size was used as it was a household characteristic and was also related to the sample weight. From experience, it was also decided to include, in addition to the household income characteristic, a dummy characteristic that identified if household income had been reported in Cycle 1 or not. As a change of address can lead to an unable-to-trace nonresponse situation we would have liked to use a change-of-address identifier. However, in some nonresponse and no contact situations it was difficult to ascertain whether the Panel Member had indeed moved. In the end a “Mover” variable, which identified whether the Panel Member had changed provinces between Cycles, was used in the analysis even though this was far from ideal because the default value would be “no”. Personal characteristics from the Health questionnaire component such as Smoker/Drinker status, Health Index Level and Mental Health/Distress Scale were not used because they were not available for almost 500 Panel Members.

The variables used are listed below. The nonresponse indicator, which was the dependent variable, had its values assigned according to the definition of nonresponse being used.

DESIGN/GEOGRAPHICAL VARIABLES

PROVINCE	The analysis was done at the provincial level
CMA	Census Metropolitan Area (0 if not a CMA)
URBAN	Urban/Rural Indicator
REJECT	Flag if the unit (household) was eligible for screening
ACFLAG	“Adult/Children” design classification for the unit

DWELLING/HOUSEHOLD CHARACTERISTICS

DWELL	Dwelling type (10 categories)
OWNER	Owner/Renter Indicator
FAMTYP	Family Type (unattached individual, single parent hhld., married couple hhld., other)
INC	Household Income Adequacy (5 levels)
INCNR	Nonresponse flag for INC
INCSRC	Main source of income (6 categories)
*HHSIZE	Household size
UND25	Indicator of members under 25 years old
KIDS	Indicator of children under 12 years old

PERSONAL CHARACTERISTICS OF PANEL MEMBER

SEX	Sex
AGE	Age in years
AGE16	Indicator if aged 16 or older
MARIT	Marital Status
FAMID	Family Identifier within household (A, B, C, ...)
RACE	White, Black, Aboriginal or Other
BORN	Place of birth (Canada, USA/Mexico, S. America/Africa, Europe/Australia, Asia)
AGIMM	Age at immigration (for immigrants)
*MOVED	Changed province indicator (see text)
*EDUC	Highest level of education (12 categories)
*STUDNT	Student Indicator
MACT	Main Activity (8 categories)
*NUMJOB	Number of jobs held last year (in Cycle 1)
RESTR	Restriction of Activity Flag
*CAUSE	Main Cause of Restriction (12 categories)
CONSUL	Number of consultations with a Medical Doctor
INHOSP	Overnight Hospital Patient Flag
*CHRONIC	Number of Chronic Conditions

* Indicates the variable was never significant when forming classes.

Figure 3 presents the variables chosen by CHAID to build nonresponse adjustment cells for Household Level Response and for the Full Longitudinal Response in each province. For reasons of confidentiality detail is not given on the individual cell sizes and response rates (some of the variables used are considered sensitive and are not on the PUMFs). However, summary information on the cell construction is given in Tables 2 and 3.

Table 2
Response Adjustment Cell Characteristics
(for Household Level Response)

Prov.	#Units	#NR	Cell Sizes			Cell % NR rates		
			min.	max.	avg.	min.	max.	avg.
Nfld.	1,082	40	354	728	541	1.4	4.8	3.7
P.E.I.	1,037	51	81	478	259	3.0	13.6	4.9
N.S.	1,085	55	46	374	217	0.7	10.9	5.1
N.B.	1,125	59	32	986	281	2.6	34.4	5.2
Que.	3,000	133	123	2363	750	1.8	12.1	4.4
Ont.	4,307	315	44	1,038	308	0.9	25.8	7.3
Man.	1,205	50	1,205	1,205	1,205	4.1	4.1	4.1
Sask.	1,168	59	37	626	167	1.6	35.3	5.1
Alta.	1,544	116	32	837	221	3.9	36.7	7.5
B.C.	1,723	149	82	678	246	5.2	29.0	8.6

The results vary by province. As expected, provinces with larger sample sizes such as Ontario, Quebec, British Columbia and Alberta yield "richer" decision trees. Cell sizes and response rates also vary considerably. In Table 2

on household-level response Manitoba has only one cell, and 88% of New Brunswick's sample is located in one cell. Likewise, in Table 3 almost all of Newfoundland's sample is placed in one of its two cells. Cell nonresponse rates approaching 40% are observed in a few provinces.

Table 3
Response Adjustment Cell Characteristics
(for Full Longitudinal Response)

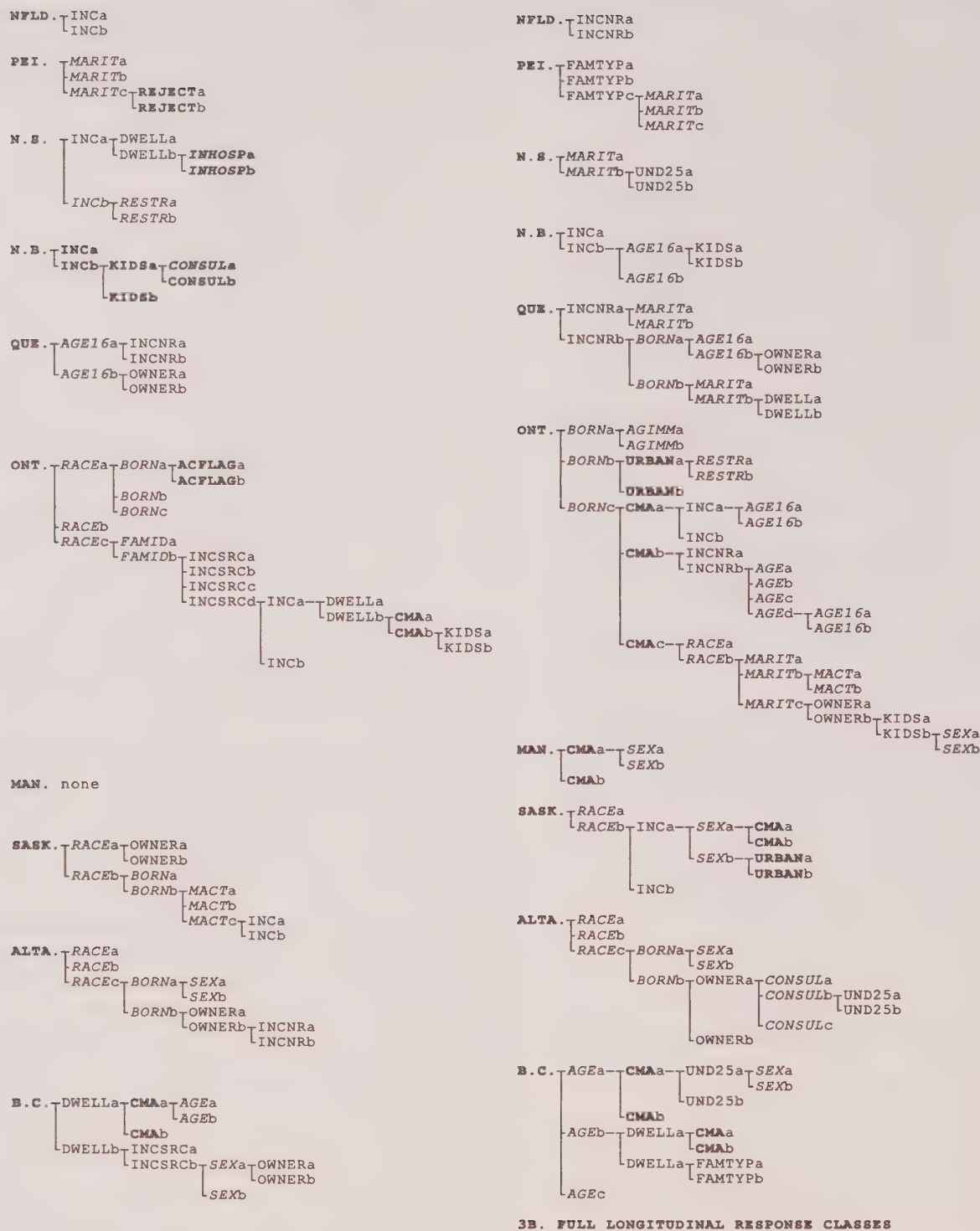
Prov.	#Units	#NR	Cell Sizes			Cell % NR rates		
			min.	max.	avg.	min.	max.	avg.
Nfld.	1,082	73	35	1,047	541	6.2	22.9	6.7
P.E.I.	1,037	80	41	453	207	4.1	26.8	7.7
N.S.	1,085	96	236	555	362	6.5	14.3	8.8
N.B.	1,125	86	59	819	281	4.8	16.8	7.6
Que.	3,000	211	42	2,202	375	2.5	37.8	7.0
Ont.	4,307	470	34	619	196	0.0	38.0	10.9
Man.	1,205	91	186	763	402	5.6	15.1	7.6
Sask.	1,168	83	90	339	195	0.0	28.9	7.1
Alta.	1,544	148	41	866	172	1.1	39.0	9.6
B.C.	1,723	192	33	408	191	4.5	37.3	11.1

Figure 3 shows a variety in the characteristics of weighting classes both between provinces and between the two types of nonresponse within provinces. In all provinces except Alberta the CHAID algorithm uses different characteristics for the two nonresponse types as early as at the first or second level of branching. A few characteristics figure prominently in the early stages of branching in many of the trees for both types of nonresponse. They are: household income adequacy level (INCNR), income non-response flag (INCNR), Race (RACE) and Place of Birth (BORN).

In Figure 3a household income and its related variables (INCNR and INCSRC), Owner/Renter status (OWNER), Race, Place of Birth and Dwelling Type (DWELL) all were used three or more times in forming weighting classes for Household Level nonresponse. It is also remarked that in five out of nine provinces a personal characteristic of the Panel Member was selected at the first stage of branching by CHAID. This supports the decision to consider personal characteristics when adjusting for household level nonresponse.

In Figure 3b for Full Longitudinal nonresponse Census Metropolitan Area (CMA), Marital Status (MARIT) and SEX, although not as important at first as Income, Race and Place of Birth, were used the most often (5 times each).

As mentioned earlier, design variables such as the rejection flag (REJECT) and the "Adult/Children" flag (ACFLAG) were considered in an attempt to incorporate the sample design in the CHAID analyses. Although these variables were selected only once each, household characteristics used by the design, such as the presence of children (KIDS) and under 25 year-olds (UND25) did get selected occasionally. Household size was not used but



(TYPES OF CHARACTERISTICS: DESIGN/GEOGRAPHICAL, DWELLING/HOUSEHOLD, PERSONAL)

Figure 3. Provincial Response Classes Obtained for Cycle 2 Nonresponse

Family Type (FAMTYP), which is related to the household size, did get selected twice.

The adjustment cells produced by CHAID were reviewed but only in rare cases were they manually altered. Within each cell, the weights of responding units were prorated to add up to the total weight for responding and nonresponding units. The magnitude of the nonresponse weight adjustments never exceeded 1.83.

4. CONCLUSION

This paper presented the strategy developed for the treatment of both longitudinal and cross-sectional nonresponse to Cycle 2 of the NPHS. The approach adopted took into account practical considerations such as the need for an easy-to-use, yet statistically valid, way of defining weight adjustment cells and the need to provide a more useful data set (by having separate adjustments for "Full" and "Partial" Longitudinal Responses) while keeping the additional effort required at a reasonable level (*e.g.*, by using weight adjustment cells for more than one purpose). Having chosen the CHAID algorithm approach rather than logistic regression allowed us more freedom in the number and choice of variables to consider: many design variables and personal variables could thus easily be considered – and some were retained. This did seem to offer some promise about the usefulness of those characteristics in the treatment of nonresponse.

On the other hand, a tight production schedule meant that some analysis that we wished to have carried out was not performed. It would have been interesting to pursue the possibilities offered by the CHAID algorithm, for example, as CHAID allows the use of a categorical response variable we could have classified sample units into three groups: live respondents, dead or out-of-scope units, and nonrespondents. We would have liked to do our own comparison of CHAID with a logistic regression approach. We could also have attempted to use Health questionnaire variables such as the Health Index or Smoker/Drinker status in defining weight adjustment cells, although their usefulness would have been reduced by the fact that they were not present for all units (they are missing in response patterns G*-GH, G*-G* and G*-**). Decisions to use the same weight adjustment cells for different types of nonresponse should be revisited. For example, could the adjustment cells built for household level response have been more suitable for the Health cross-sectional nonresponse? An attempt to compare the efficiency of various nonresponse adjustment strategies would involve evaluating their impact on the variance of estimators. We could also evaluate the impact of our Cycle 2 nonresponse adjustment on the nonresponse bias by using the Cycle 1 data available for all panel members. Estimates using the full sample would be compared to nonresponse-adjusted estimates generated from the responding units.

Cycle 3 itself will present new problems. A global sample "top-up" is planned in that year, which will have an impact on our cross-sectional estimation strategy and therefore on the treatment of nonresponse. As longitudinal nonresponse is increasing we will have to consider side effects of the weighting adjustment such as the possible creation of outlier weights. Providing sets of weights for different types of longitudinal analyses will become cumbersome as the number of "partial" response patterns will increase. How many patterns can reasonably be treated, and which ones? The choice of additional information, such as Mover status, for the treatment of nonresponse should be reconsidered. Some imputation for nonresponse will likely be used in Cycle 3: the question is how to reconcile imputation with the weight adjustment approach to nonresponse. As can be seen, a lot of work remains to be done for the NPHS. One hopes that we will have time to investigate many of those issues before Cycle 3 processing is finished.

ACKNOWLEDGEMENTS

The authors would like to thank Douglas Yeo and the referees for their helpful comments.

REFERENCES

- ANGOSS SOFTWARE (1995). *Knowledge SEEKER IV for Windows – User's Guide*. ANGOSS Software International Limited.
- CATLIN, G., and WILL, P. (1992). The National Population Health Survey: Highlights of initial developments. *Health Reports*, Statistics Canada, 4, 313-319.
- CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., and RUBIN, B.R. (1992). Projecting from advance data using propensity modelling: an application to income and tax statistics. *Journal of Business & Economic Statistics*, 10, 117-131.
- DAVID, M.H., LITTLE, R., SAMUHEL, M., and TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.
- DUFOUR, J., GAGNON, F., MORIN, Y., RENAUD, M., and SÄRNDAL, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. To appear.
- ERNST, L. (1989). Weighting issues for household and family estimates. In *Panel Surveys*, (Eds. D. Kasprzyk, G. Duncan, G. Kalton and M.P. Singh). New York: John Wiley and Sons, 135-159.
- GRONDIN, C. (1996). Pondération longitudinale – Première vague de l'EDTR. Internal note, Social Survey Methods Division, Statistics Canada.

- INSTITUTE FOR SOCIAL RESEARCH (1979). A Panel Study of Income Dynamics: Procedures and Tape Codes – 1978 Interviewing Year – Wave XI – A Supplement. The University of Michigan.
- KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.
- KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.
- LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.
- LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 54, 137-157.
- OH, H.L., and SCHEUREN, F. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, (Eds. W.G. Madow, I. Olkin and D. Rubin). New York: Academic Press, 143-184.
- RIZZO, L., KALTON, G., and BRICK, J.M. (1996). A comparison of some weighting adjustment methods for panel nonresponse. *Survey Methodology*, 22, 43-53.
- ROSENBAUM, P.R., and RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.
- SKINNER, C.J., and RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 433, 349-356.
- STATISTICS CANADA (1995). National Population Health Survey 1994-95 public use microdata file. Catalogue No. 82F0001XCB.
- STUKEL, D.M., MOHL, C.A., and TAMBAY, J.L. (1997). Weighting for cycle two of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 111-116.
- TAMBAY, J.L., and CATLIN, G. (1995). Sample design of the National Population Health Survey. *Health Reports*, Statistics Canada, 7, 29-38.
- THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistik Tidskrift*, 11, 278-283.

Estimates of the Errors in Classification in the Labour Force Survey and Their Effect on the Reported Unemployment Rate

MICHAEL D. SINCLAIR and JOSEPH L. GASTWIRTH¹

ABSTRACT

This paper studies response errors in the Current Population Survey of the U.S. Bureau of the Census and assesses their impact on the unemployment rates published by the Bureau of Labour Statistics. The measurement of these error rates is obtained from reinterview data, using an extension of the Hui and Walter (1980) procedure for the evaluation of diagnostic tests. Unlike prior studies which assumed that the reconciled reinterview yields the true status, the method estimates the error rates in both interviews. Using these estimated error rates, we show that the misclassification in the original survey creates a cyclical effect on the reported estimated unemployment rates. In particular, the degree of underestimation increases when true unemployment is high. As there was insufficient data to distinguish between a model assuming that the misclassification rates are the same throughout the business cycle, and one that allows the error rates to differ in periods of low, moderate and high unemployment, our findings should be regarded as preliminary. Nonetheless, they indicated that the relationship between the models used to assess the accuracy of diagnostic tests, and those measuring misclassification rates of survey data, deserves further study.

KEY WORDS: Misclassification errors; Unemployment rates; Diagnostic tests; Reconciliation; Reinterview surveys; Response errors.

1. INTRODUCTION

Several articles, Poterba and Summers (1986 and 1995) and Abowd and Zellner (1985) used the data from the U.S. Bureau of the Census' reinterview program to estimate the misclassification rates of the Current Population Survey (CPS) and assessed their impact on estimates of labour market transition rates. The estimated misclassification rates were based on the assumption, that a particular reinterview method, reconciliation, yields the "truth." Biemer and Forsman (1992), Forsman and Schreiner (1991) and unpublished research of the U.S. Bureau of the Census (1963), have questioned this assumption. The purpose of this paper, is to provide estimates of the misclassification rates, from response errors in all interviews and reinterviews and to explore their impact on the reported unemployment rates. In contrast to the earlier papers that were concerned with gross flow, we emphasize the accuracy of the labour force estimates themselves. Our approach is based on extending the Hui and Walter (1980) paradigm, for estimating error rates of medical diagnostic tests to trinomial classifications. An advantage of this method is that, no single interview needs to be considered as perfect.

Under certain assumptions, Hui and Walter (1980) developed a method for estimating the error rates associated with a new diagnostic screening test, using a confirmatory test with an unknown low error rate. By treating the reinterview as the confirmatory test, and the original survey as the screening test, this methodology can be used to estimate the error rates in the original survey, and the reinterview and the prevalence rates of the trait screened for. The

Hui and Walter (1980) method requires two subpopulations with different prevalence rates of the characteristic. While the two tests may have different error rates, the error rates for each test are assumed equal in the two subpopulations. Furthermore, the model (described in more detail in the appendix) assumes that the errors from the two tests conditioned on the subject's true status, are independent.

The Hui and Walter method was developed for dichotomous test outcomes, and was adapted by Sinclair and Gastwirth (1996) to study misclassification of labour force participation rates. Here, we extend the approach to account for three classifications: unemployed, employed and not in the labour force (NLF), and assess the effect of the misclassification on the reported unemployment rates. The basic model is presented in section two. The reinterview program data, to which the model will be fitted, are described in section three. The resulting error rates are given in section four, along with the "adjusted" unemployment rates, which account for the estimated classification errors. In addition, a measure of accuracy, the predictive value, used in the medical screening literature, is applied to the unemployment rate in section four. It shows that the probability an individual classified as unemployed in the CPS is actually unemployed, varies with the true level of unemployment.

2. THE DATA AND THE MODEL

Labour force reinterview data consists of trinomial responses from both the original survey and a subsequent reinterview. This data for a given subpopulation and year,

¹ Michael D. Sinclair, Senior Statistician, Mathematical Policy Research, 101 Morgan Lane, Plainsboro, N.J. 08536, U.S.A.; Joseph L. Gastwirth, Professor of Statistics and Economics, George Washington University, 2201 G Street Rm. 315, Washington, D.C. 20052, U.S.A.

is summarized in a 3×3 table, where the observed frequency counts of persons in the table, is denoted by, n_{ygi} . With this notation:

- y denotes the year;
- g denotes subpopulation membership, $g = 1$ or 2 ;
- i denotes the subject's classification by the original survey, $i = 1$ for unemployed, $i = 2$ for employed and $i = 3$ for NLF; and
- j denotes the same subject's classification by the reinterview, $j = 1, 2$ and 3 .

We denote the true prevalence rate for each labour force status, $i = 1, 2$ and 3 , by π_{ygi} , for subpopulation g and year y . Throughout this paper, we will use the term prevalence rate, to refer to the proportion of persons in one of the three labour force categories (e.g., π_{yg1}). Note that the fraction, π_{yg3} of the population in the NLF category equals $(1 - \pi_{yg1} - \pi_{yg2})$, and that the true unemployment rate in year y for subpopulation g , is equal to π_{yg1} divided by $(\pi_{yg1} + \pi_{yg2})$.

Each classification rate, β_{ygrij} , is defined as the probability that the r -th data collection process, $r = 1$ for the original survey, and $r = 2$ for the reinterview, will classify a person in year y from subpopulation g , to be in category i , $i = 1, 2$ and 3 when the true status of the individual is category j . For example, β_{11131} denotes the probability that in the first year ($y = 1$), a person from the first subpopulation ($g = 1$), was classified by the original survey ($r = 1$) as NLF ($i = 3$) when the person's true status is unemployed ($j = 1$). The classification rates can be divided into two groups, corresponding to those associated with a correct classification, and those associated with an erroneous classification. For each y, g and r , the probability that survey method r , classifies a truly unemployed person in year y from subpopulation g correctly as unemployed, is equal to $\beta_{ygr11} = (1 - \beta_{ygr21} - \beta_{ygr31})$. The corresponding probabilities for employed and NLF are respectively, $\beta_{ygr22} = (1 - \beta_{ygr12} - \beta_{ygr32})$, and $\beta_{ygr33} = (1 - \beta_{ygr13} - \beta_{ygr23})$. With conditional independence of the original survey and the reinterview classification rates, the expected observed frequencies, as expressed in terms of the given notation, for each of the nine cells associated with a particular year y and subpopulation g are:

$$E(n_{yg11}) = n_{yg..} (\pi_{yg1} (1 - \beta_{yg121} - \beta_{yg131}) (1 - \beta_{yg221} - \beta_{yg231}) + \pi_{yg2} \beta_{yg112} \beta_{yg212} + (1 - \pi_{yg1} - \pi_{yg2}) \beta_{yg113} \beta_{yg213})$$

$$E(n_{yg12}) = n_{yg..} (\pi_{yg1} (1 - \beta_{yg121} - \beta_{yg131}) \beta_{yg221} + \pi_{yg2} \beta_{yg112} * (1 - \beta_{yg212} - \beta_{yg232}) + (1 - \pi_{yg1} - \pi_{yg2}) \beta_{yg113} \beta_{yg223})$$

$$E(n_{yg13}) = n_{yg..} (\pi_{yg1} (1 - \beta_{yg121} - \beta_{yg131}) \beta_{yg231} + \pi_{yg2} \beta_{yg112} \beta_{yg232} + (1 - \pi_{yg1} - \pi_{yg2}) \beta_{yg113} (1 - \beta_{yg213} - \beta_{yg223}))$$

$$E(n_{yg21}) = n_{yg..} (\pi_{yg1} \beta_{yg121} (1 - \beta_{yg221} - \beta_{yg231}) + \pi_{yg2} (1 - \beta_{yg112} - \beta_{yg132}) \beta_{yg212} + (1 - \pi_{yg1} - \pi_{yg2}) \beta_{yg123} \beta_{yg213})$$

$$E(n_{yg22}) = n_{yg..} (\pi_{yg1} \beta_{yg121} \beta_{yg221} + \pi_{yg2} (1 - \beta_{yg112} - \beta_{yg132}) * (1 - \beta_{yg212} - \beta_{yg232}) + (1 - \pi_{yg1} - \pi_{yg2}) \beta_{yg123} \beta_{yg223})$$

$$E(n_{yg23}) = n_{yg..} (\pi_{yg1} \beta_{yg121} \beta_{yg231} + \pi_{yg2} (1 - \beta_{yg112} - \beta_{yg132}) \beta_{yg232} + (1 - \pi_{yg1} - \pi_{yg2}) \beta_{yg123} (1 - \beta_{yg213} - \beta_{yg223}))$$

$$E(n_{yg31}) = n_{yg..} (\pi_{yg1} \beta_{yg131} (1 - \beta_{yg221} - \beta_{yg231}) + \pi_{yg2} \beta_{yg132} \beta_{yg212} + (1 - \pi_{yg1} - \pi_{yg2}) (1 - \beta_{yg123} - \beta_{yg113}) \beta_{yg213})$$

$$E(n_{yg32}) = n_{yg..} (\pi_{yg1} \beta_{yg131} \beta_{yg221} + \pi_{yg2} \beta_{yg132} (1 - \beta_{yg212} - \beta_{yg232}) + (1 - \pi_{yg1} - \pi_{yg2}) (1 - \beta_{yg123} - \beta_{yg113}) \beta_{yg223})$$

$$E(n_{yg33}) = n_{yg..} (\pi_{yg1} \beta_{yg131} \beta_{yg231} + \pi_{yg2} \beta_{yg132} \beta_{yg232} + (1 - \pi_{yg1} - \pi_{yg2}) (1 - \beta_{yg123} - \beta_{yg113}) (1 - \beta_{yg213} - \beta_{yg223})),$$

where, the total sample size for year y and subpopulation g is denoted by $n_{yg..}$.

The model has 14 parameters (six error rates for the original survey, $r = 1$, six error rates for the reinterview, $r = 2$, and two unique prevalence rates) for each subpopulation and year. On the other hand, the 3×3 table for a given year and subpopulation has only 8 independent frequencies, or degrees of freedom. As a result, the model is overparameterized and the number of parameters must be reduced for estimation purposes. The Hui and Walter paradigm enables us to accomplish this.

3. APPLICATION OF THE MODEL AND THE CPS REINTERVIEW PROGRAM

The U.S. Bureau of the Census' Current Population Survey Reinterview Program (U.S. Bureau of the Census 1963) is conducted approximately two weeks after the initial survey, to measure response errors, and to evaluate interviewer performance. The sample design for the reinterview, consists of the self-weighting random sample of households (Levy and Lemeshow 1980) among the selected interviewer assignments. The sample size is about 1/18 of the monthly CPS sample of 50,000 to 60,000 household interviews. Two reinterview procedures are conducted. Three-fourths to four-fifths of the sample cases participate in a response-bias study. Here, an initial reinterview is conducted and after this interview is

completed, the reinterviewer reconciles disagreements with the respondent, between the original and the initial reinterview responses. Hence, in the response-bias study, up to two reinterview responses may be obtained from each subject; the first unreconciled reinterview response and a reconciled reinterview response. The remaining one-fifth to one-fourth of the sample households receive a reinterview without reconciliation.

In the response bias study, the reinterviewer is instructed not to look at the original survey responses until the initial reinterview is completed. Forsman and Schreiner (1991) and Schreiner (1980) suggested that the reinterviewers may change the initial reinterview responses to match the original response, as they observed that the rate of disagreement between the original responses and the initial reinterview responses were greater in the unreconciled sample. Sinclair (1994) and Sinclair and Gastwirth (1996) showed that these differences were statistically significant. As a result, the reconciliation process creates a correlation between the original and unreconciled reinterview responses, in the reconciled sample. Hence, we decided to limit our analysis to the original and unreconciled reinterview data from the unreconciled study sample. For the purposes of this study, we will assume that in the unreconciled sample, the errors from the original survey and the unreconciled reinterview conditioned on the respondent's true status, are independent.

To apply the Hui and Walter approach, one needs two subpopulations with different prevalence rates. As males and females are known to have different labour force participation rates, we use them. We also need to assume, that the classification error rates are equal in the two subpopulations, males and females, *i.e.*, $\beta_{y1rij} = \beta_{y2rij}$. At this stage, we assume that the classification error rates for the original survey and the unreconciled reinterview, may be different, and that they may differ by year. With this reduction, for the two subpopulations, in a given year, we now have a total of 12 error rate parameters and 4 prevalence rates, yielding 16 parameters. Since two 3×3 tables contain a total of 16 degrees of freedom, estimation is possible. In this paper, we have analyzed the CPS unreconciled reinterview sample data for the period 1981 through 1990. Complete yearly data for 1987 as well as more recent data, were not available from the U.S. Bureau of the Census.

The CPS estimates of the unemployment rate are published regularly by the Bureau of Labour Statistics (BLS) (see Bureau of Labour Statistics 1992). Since the reinterview is a sub-sample of the full CPS sample, the original survey estimates of the unemployment rate from the reinterview sample, will differ from the BLS published results. Data processing procedures are used on the full sample CPS, that are not applied to the reinterview data. For example, the full CPS sample is weighted, based on the sample selection probabilities, and nonresponse adjustment factors are applied to the data. Given these differences, the estimated prevalences from our model, based solely on the reinterview data,

are not directly comparable to the BLS reported values. We have used the CPS reinterview data, primarily to estimate the error rates in the original survey. Furthermore, we have treated the unreconciled reinterview data as a simple random sample of the population, for analysis and hypothesis testing purposes, throughout this paper. Using these error rate estimates, we estimate adjusted Bureau of Labour Statistics (BLS) unemployment rates, where the term adjusted, means that the reported values have been modified to account for the misclassification in the survey. The formula for estimating the true unemployment rate as a function of the reported BLS prevalences from the full CPS sample, and the estimated classification error rates as obtained from the unreconciled reinterview data, is given in the appendix.

4. DATA ANALYSIS AND RESULTS

The first step in preparing our final estimates, was to obtain the parameter estimates, for each of nine yearly data tables, using the SAS NLIN procedure with the Gauss-Newton weighted least squares method. As the reinterview procedures remained constant during the period, we decided to test the hypothesis, that each of the error rates remained equal across the years studied, *i.e.*, $\beta_{ygrij} = \beta_{y'grij}$ for all years $y \neq y'$. In conjunction with the basic assumption, that the error rates for males and females are equal, *i.e.*, $\beta_{y1rij} = \beta_{y2rij}$, this implies, $\beta_{ygrij} = \beta_{y'g'rij}$ for all $y \neq y'$ and $g \neq g'$.

From the two sets of results, we conducted a likelihood ratio test under the assumption, that the reinterview sample is a simple random sample of the population, to test the assumption that each of the error rates was the same for all years. The likelihood ratio statistic, $-2 \log \lambda$ with 96 degrees of freedom (144 parameters in the full model less 48 parameters in the reduced model) yielded a value of 84.06 with a p -value of 0.8027. Hence, the data is consistent with the reduced model, enabling us to use the reduced model estimates and to simplify the notation. We will now use β_{rij} to denote β_{ygrij} for all g and y .

The estimated error rates for the original survey and for the unreconciled reinterview, are presented in Tables 1 and 2, respectively, with their estimated standard errors. The estimated reinterview error rates in Table 2, are similar to corresponding error rate estimates for the original survey. This similarity indicates that the U.S. Bureau of the Census unreconciled reinterview serves as an effective replication. The error rate estimates show that the CPS survey procedures are able to classify the employed, and those not in the labour force, quite accurately. On the other hand, these procedures do not perform well for classifying the unemployed, as the proportion of truly unemployed persons who are classified as unemployed, $(1 - \beta_{121} - \beta_{131})$, is only 0.8397.

For comparative purposes we conducted an analysis of the 75% sample reconciled reinterview data, for the same

1981-1990 period, under the assumption that the reconciled responses were error-free. We created a 3×3 table for the number of persons classified by the original interview, in each labour force category, by the number of persons classified by the reconciled reinterview, in each labour force category. The data is given in Table 3. The table frequencies report aggregate data, by year and sex, so that the error rates derived from this table, are comparable to our model. Using the column status, as the true status, one computes an estimate of the error rates. For example, the estimate of β_{121} , the probability that an unemployed person will be classified in the original survey as employed, is $332/17,681 = 0.0188$. These error rates are presented in Table 1, to illustrate how the estimated error rates from our method, based on the unreconciled data, differ from those relying on the assumption that the reconciled reinterview is perfect.

Table 1 also presents the estimates of the original survey error rates, as obtained by Poterba and Summers (1986), using reinterview data (combined for both sexes) for the first half of 1981. The Poterba and Summers' method uses both the data from the unreconciled and reconciled samples to estimate the error rates. These authors assume that in the reconciled sample, the interviewers use the original survey data provided, to influence the initial reinterview response. As a result, they assume that a reconciled value is only obtained for a portion of persons, that should have had a

discrepancy between the original survey and the initial reinterview. When a reconciled value is obtained, Poterba and Summers assume that the reconciled data is error-free. With these assumptions, they use the unreconciled sample to estimate the incidence of the error, and the reconciled data to provide the information on the true labour force status. In summary, both the Poterba and Summers method, and the reconciled reinterview estimates, rely on the reconciled reinterview data being perfect.

Table 4 presents the reported BLS yearly unemployment rates among those in the labour force, for males and females combined, in comparison to the estimated adjusted unemployment rates based on: (1) our error rate estimates, (2) Poterba and Summers (1986) error rates, and (3) error rates assuming the reconciled reinterview is perfect. If the results in Table 4, are sorted by the value of the BLS reported unemployment rate, an apparent trend is observed in the bias in the original CPS estimates. Figure 1 shows that the reported values, tend to overestimate the actual unemployment rate of persons in the labour force in low unemployment years (1989, 1988 and 1990), and to underestimate the unemployment rate in high unemployment years (1982-1983). Furthermore, the bias associated with our method is shifted upward from the two other approaches. All three methods indicate cyclical effect, the smallest of which is obtained when the reconciled reinterview is assumed perfect.

Table 1
Estimated Error Rates in the Original CPS Estimates

Error Rate Parameter	Description		Estimated Value β_{1ij}			Estimated Standard Error
	Classified as	True Status	Our Method	P&S (1986)	Recon. Reint. Perfect	Our Method
β_{121}	Employed	Unemployed	0.0407	0.0378	0.0188	0.01892
β_{131}	NLF	Unemployed	0.1196	0.1146	0.0838	0.01463
β_{112}	Unemployed	Employed	0.0049	0.0054	0.0017	0.00124
β_{132}	NLF	Employed	0.0100	0.0172	0.0098	0.00154
β_{113}	Unemployed	NLF	0.0110	0.0064	0.0034	0.00155
β_{123}	Employed	NLF	0.0205	0.0116	0.0053	0.00247

Table 2
Estimated Error Rates in the Unreconciled Reinterview CPS Estimates

Error Rate Parameter	Description		Estimated Value	Estimated Standard Error
	Classified as	True Status	Our Method β_{2ij}	
β_{221}	Employed	Unemployed	0.0333	0.01772
β_{231}	NLF	Unemployed	0.1128	0.01360
β_{212}	Unemployed	Employed	0.0057	0.00135
β_{232}	NLF	Employed	0.0145	0.00160
β_{213}	Unemployed	NLF	0.0157	0.00171
β_{223}	Employed	NLF	0.0248	0.00238

Table 3
Cross-tabulation of the Aggregated 1981-1990 Original/Reconciled Reinterview Responses
75% Reconciled CPS Reinterview Data

Survey Result	Reconciled Reinterview			
Original CPS	Unemployed	Employed	NLF	Total
Unemployed	15,868	372	480	16,720
Employed	332	213,987	744	215,063
NLF	1,481	2,123	138,077	141,681
Total	17,681	215,482	139,301	373,464

Table 4
Implications of the Error Rate Estimates

Year y	BLS Reported Unemployment Rate UE_y^{BLS}	Prob. Unemp. Given Classified Unemp.	Adjusted Estimate of BLS Reported Unemployment Rate AUE_y^{BLS}			Difference in Reported vs. Adjusted	Estimated Standard Error in Difference
			Our Method	Poterba and Summers (1986)	Reconciled Data (1981-1990) Perfect		
1990	5.44%	.8135	5.27%	5.36%	5.63%	0.17%	.27%
1989	5.20%	.8052	4.99%	5.09%	5.37%	0.21%	.26%
1988	5.43%	.8113	5.25%	5.35%	5.62%	0.18%	.27%
1986	6.89%	.8503	6.97%	7.04%	7.22%	-0.08%	.33%
1985	7.09%	.8531	7.20%	7.27%	7.44%	-0.11%	.34%
1984	7.41%	.8581	7.56%	7.63%	7.79%	-0.15%	.36%
1983	9.47%	.8894	9.99%	10.00%	10.04%	-0.52%	.48%
1982	9.54%	.8902	10.08%	10.09%	10.12%	-0.54%	.49%
1981	7.50%	.8581	7.66%	7.72%	7.88%	-0.16%	.36%

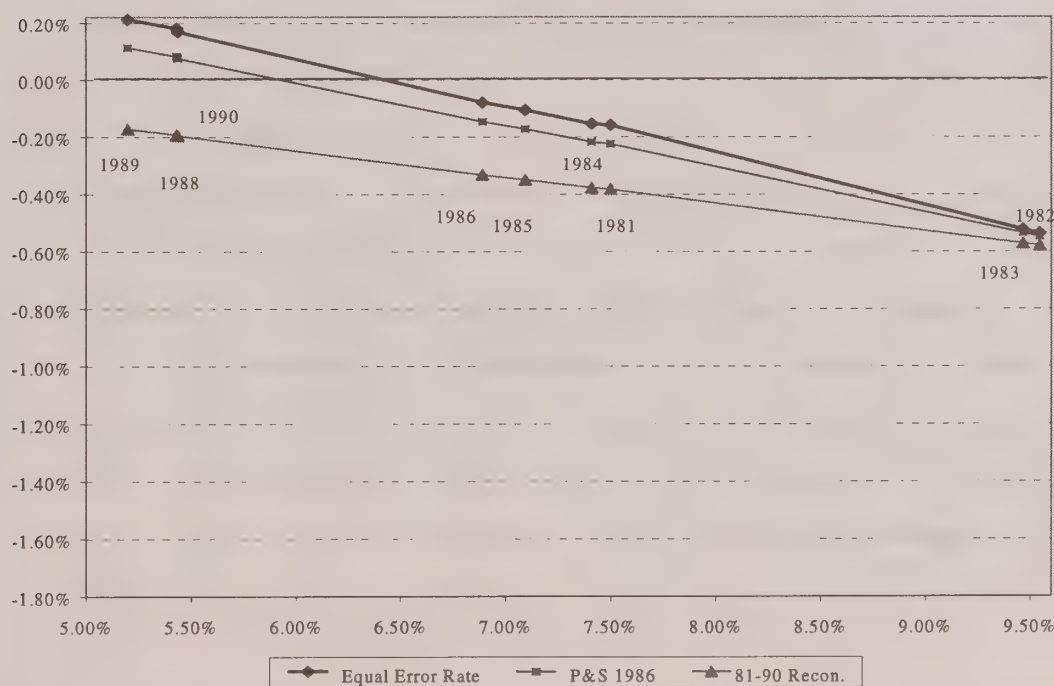


Figure 1. A Comparison of the Bias in the Reported Unemployment Rates as Computed Using Three Methods

In the screening test literature (Gastwirth 1987), the fraction of positive classifications which are correct, called the predictive value of a positive test, is known to vary directly with the prevalence of the characteristic. This is why, quite accurate diagnostic tests can have unacceptably high misclassification rates when populations with a low prevalence of a disease, are screened with them. The analog of this measure in our context, is the proportion of individuals classified as unemployed who are truly unemployed. This proportion is given in the third column of Table 4. Even though the range of reported unemployment rates is fairly narrow, a similar relationship with the unemployment rate can be seen.

While the results of the likelihood ratio test indicated, that the error rates were constant throughout the period, the referees suggested a further analysis to explore this assumption. We divided each of the nine survey years into three groups, according to the year's reported unemployment rate. Survey years, 1990, 1989 and 1988 were classified as having low unemployment, with reported rates from 5.20% to 5.44%. Similarly, survey years 1982 and 1983 were classified as having high unemployment, with reported rates of 9.54% and 9.47%, respectively. The remaining years with rates ranging from 6.89% to 7.5%, were classified as having moderate unemployment rates. With this three group structure, we developed an alternative model that assumed that the error rates were constant within each of the three rate size groups, but allowed each of these groups to have

different error rates. The estimated error rates for the original interview are presented in Table 5. The error rates from Table 1, using the equal error rate model, are presented for comparative purposes.

We conducted a likelihood ratio test, to test the assumption that each of the error rates was the same, within each of these three groups, in comparison to the initial nine year model. The likelihood ratio statistic, $-2 \log \lambda$ with 72 degrees of freedom (144 parameters in the full model less 72 parameters in the three-group model), yielded a value of 69.25 with a p -value of 0.5697.

In general, the error rate estimates for the three unemployment rate classes, appear to be similar. Because the standard errors of the estimated error rates are quite large, a formal homogeneity test would have insufficient power to detect any variation in an error rate over the three periods.

To assess the sensitivity of the adjusted unemployment rate estimates in Table 4, we recomputed them using the error rates from the three-group model. The results are given in Table 6, which also provides the standard error of the unemployment rate estimates, ranging from a low of about 1.4% to a high of about 2.6%.

Figure 2 presents a graph of the bias in the unemployment using the three group model, and for comparison, the original equal error rate model. The results in Figure 2 are quite interesting. While the cyclical effect is still apparent, the estimated bias is shifted downward and shows a consistent negative bias throughout the business cycle.

Table 5
Error Rates in the Original CPS Data Estimated for Three Unemployment Rate Classes

Error Rate Parameter	Description		Error Rate Estimates							
			Model in Table 1 Assumes Constant Error Rates Across Years		Estimates Using Three Group Model					
	Classified as	True Status			Low Years 1990, 1989, & 1988		Moderate Years 1981, 1984-1986		High Years 1982, 1983	
			Est.	STE	Est.	STE	Est.	STE	Est.	STE
β_{121}	Employed	Unemployed	0.0407	0.0189	0.0635	0.1061	0.1113	0.1258	0.0974	0.0717
β_{131}	NLF	Unemployed	0.1196	0.0146	0.1680	0.0538	0.1000	0.0246	0.1084	0.0221
β_{112}	Unemployed	Employed	0.0049	0.0012	0.0000	0.0047	0.0000	0.0098	0.0000	0.0069
β_{132}	NLF	Employed	0.0100	0.0015	0.0080	0.0038	0.0096	0.0025	0.0096	0.0031
β_{113}	Unemployed	NLF	0.0110	0.0015	0.0096	0.0040	0.0109	0.0024	0.0103	0.0029
β_{123}	Employed	NLF	0.0205	0.0025	0.0187	0.0065	0.0202	0.0034	0.0227	0.0044

Table 6
Implications of the Error Rate Estimates Using Three Group Model

Year y	BLS Reported Unemploy- ment Rate	Prob Unemp. Given Classified Unemp. Three Group Model	Adjusted Estimate of BLS Reported Unemployment Rate		Difference in Reported vs. Adjusted		
			Original Equal Error Rate Model	Three Group Model	Original Equal Error Rate Model	Three Group Model	Estimate Standard Error of the Difference Three Group Method
1990	5.44%	0.9124	5.27%	6.43%	0.17%	-0.99%	1.40%
1989	5.20%	0.9088	4.99%	6.12%	0.21%	-0.93%	1.35%
1988	5.43%	0.9105	5.25%	6.41%	0.18%	-0.98%	1.41%
1986	6.89%	0.9170	6.97%	8.01%	-0.08%	-1.12%	2.35%
1985	7.09%	0.9178	7.20%	8.25%	-0.11%	-1.16%	2.42%
1984	7.41%	0.9199	7.56%	8.64%	-0.15%	-1.23%	2.53%
1983	9.47%	0.9400	9.99%	11.18%	-0.52%	-1.71%	2.05%
1982	9.54%	0.9404	10.08%	11.27%	-0.54%	-1.73%	2.08%
1981	7.50%	0.9191	7.66%	8.74%	-0.16%	-1.24%	2.56%

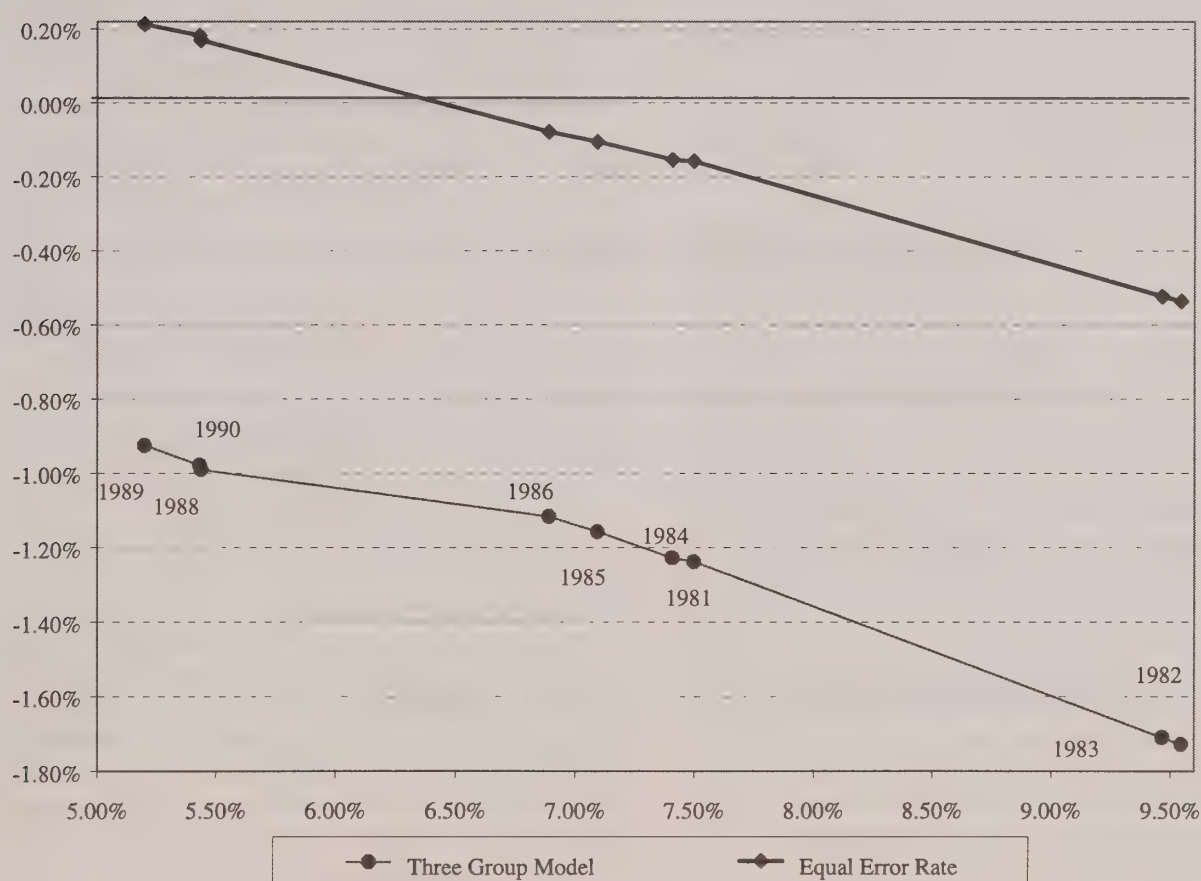


Figure 2. A Comparison of the Bias in the Reported Unemployment Rates as Computed Using the Equal Error Rate Model and the Three Group Model

5. IMPLICATIONS OF THE ADJUSTED ESTIMATES

The results in Figure 1 and 2 show that, all methods for adjusting the unemployment rate for misclassification error, indicate that the degree of bias in the reported rate varies over the business cycle. Given the differences in the estimated bias yielded by the two approaches, it is difficult to determine the magnitude of the bias. Unfortunately, the estimates are sensitive to the model specification, due to the small unreconciled reinterview sample size. This is reflected in the large standard errors of the estimated error rates, and consequently, the estimated bias.

Our approach using the assumption that the error rates remained constant throughout, suggests that bias in the survey estimates is small in years when the unemployment rate is between 5.5% and 7.5%. With this model, the reported unemployment rate appears to be unbiased when the true unemployment rate is around 6.3%, and yields an underestimate when the true rate is above this level, and an overestimate when the true rate is below it. The underestimation bias becomes quite noticeable when unemployment reaches 9%, while the overestimation bias could be meaningful when unemployment is less than 5%.

Using the three-group model results, implies that the reported unemployment rates are underestimates. If the finding is accurate, these results show that the bias in low unemployment years is still about -0.7%, but can be as high as -1.7% in high unemployment years. This contrasts the results obtained from the equal error rate model.

The fact that both the magnitude and direction of the bias in the reported unemployment rate change over the business cycle, may affect the use of that rate in studies of the "natural rate" of unemployment, and the trade-off between inflation and unemployment. Specifically, our results indicate that the range of the true unemployment rate over the business cycle, is larger than the range of the reported rate (see Table 4). Hughes and Perlman (1984) survey the literature on the "natural rate" of unemployment, and the trade-off between inflation and unemployment, as well as the role of search theory in explaining why unemployment is not that low at "full" employment. McKenna (1985) provides a more advanced treatment of job search theory, and its relationship to the duration of unemployment, and the degree to which unemployment is voluntary. Resolving the issue of which model underlies the misclassification error rates in the CPS survey, has important economic implications. If the equal error rate model were correct, in periods of low unemployment, the reported rate would be a slight overestimate. Hence, there would be less true unemployment to explain, by job search and related theories. On the other hand, if the three group model is the correct one, then even at low levels of reported unemployment, there are more persons really unemployed.

6. DISCUSSION

In this paper, we have presented an alternative method for estimating the error rates in the CPS survey. Our study differs from prior work, as we follow the Hui and Walter (1980) approach to estimate the error rates, by assuming that males and females will have the same error rates, and that the errors in the original survey are independent of those in the unreconciled reinterview. While the errors could be slightly correlated, the assumption of independence is standard in data analysis of this type, (see Bailar 1968, Chua and Fuller 1987, and Singh and Rao 1995). A discussion of the bias in the H&W method with dependent errors is given in Vacek (1985). As for the equal error rate assumption, several of the authors cited in this paper (e.g., Poterba and Summers 1986), have noted minor to moderate differences in the error rates between males and females, under the assumption that the reconciled reinterview is perfect. However, this assumption has been questioned. For example, consider the estimate of β_{121} , the probability that an unemployed person, will be classified in the original survey as employed. From Table 3, we estimate this value under the assumption that the reconciled reinterview is unbiased, by dividing n_{21} , divided by $n_{.1}$ ($332/17,681 = 0.0188$), where n_{ij} is defined previously, with j now corresponding to the classification status in the reconciled reinterview. Using the expected value of these two frequencies from section 2, we can write an expression for the expectation of the estimate in large samples as follows:

$$\begin{aligned}
 E(n_{21}/n_{.1}) &= \frac{\pi_1\beta_{121}(1-\beta_{221}-\beta_{231})+\pi_2(1-\beta_{112}-\beta_{132})\beta_{212}+(1-\pi_1-\pi_2)\beta_{123}\beta_{213}}{\pi_1(1-\beta_{221}-\beta_{231})+\pi_2\beta_{212}+(1-\pi_1-\pi_2)\beta_{213}} \\
 &= \beta_{121} + \beta_{121} \left[\frac{\pi_1(1-\beta_{221}-\beta_{231})}{\pi_1(1-\beta_{221}-\beta_{231}) + \pi_2\beta_{212} + (1-\pi_1-\pi_2)\beta_{213}} - 1 \right] \\
 &\quad + \left[\frac{\pi_2(1-\beta_{112}-\beta_{132})\beta_{212}+(1-\pi_1-\pi_2)\beta_{123}\beta_{213}}{\pi_1(1-\beta_{221}-\beta_{231}) + \pi_2\beta_{212} + (1-\pi_1-\pi_2)\beta_{213}} \right]. \tag{1}
 \end{aligned}$$

From (1) it follows that, if the reconciled reinterview error rates, β_{2ij} are equal to zero, that this estimator is unbiased. However, if the reconciled reinterview is not perfect, then the bias in the estimator depends on the prevalence rates in the population studied. As a result, if the actual original survey error rates are in fact equal in the two subpopulations studied, and the reconciled survey classifications are not perfect, the estimated original survey error rates for the two populations will differ. Therefore, one cannot use the similarities or differences in the estimated error rates for males and females from earlier

papers, to justify or to contradict the assumptions used here.

We have also conducted a sensitivity analysis of the Hui and Walter (1980) method for dichotomous responses (Sinclair 1994), that indicates that the procedure is sensitive to a violation in the equal error rate assumption, in some circumstances, but the procedure is quite robust in others. Further research is needed to develop reinterview procedures and analytical techniques, to relax the restrictive assumptions currently required in the analysis of the reinterview data.

It should be noted that Chua and Fuller (1987) also obtained estimates of the 3-outcome classification errors in the 1977-1980 CPS 25% sample reinterview data. Analogous to our results, their study found that the largest error rates were associated with classifying the truly unemployed. Poterba and Summers (1995) and Singh and Rao (1995) also found this group to be the hardest to classify. Because all models examined, indicated that the overall misclassification rate of an unemployed individual is around 20%, future reinterviews might focus on understanding why these rates are so high. Hopefully, this will lead to an improved survey.

A potential use of the "adjusted" estimates in Table 4, is in a sensitivity analysis of the literature (*e.g.*, Abowd and Zellner 1985; Poterba and Summers 1995) on gross flows, and labour market dynamics, which assumed that the reconciled interview was perfect. This is equivalent to their adoption of the estimates in the next to the last column of Table 3. Similarly, estimates of the classification errors may be incorporated in procedures, for estimating probit and logit models with misclassified response variables (Hausman and Morton 1994), and in the development of formal statistical procedures for survey data (Rao and Thomas 1991). It should be emphasized, that all the estimates adjusting for misclassification, are still in the research phase, and that the error rates are not yet estimated with sufficient accuracy, to adjust the regular survey data, especially as a new questionnaire and new interviewing procedures were introduced as of January 1994 (Bureau of Labour Statistics 1993).

ACKNOWLEDGEMENTS

The authors wish to thank Irv Schreiner of the U.S. Bureau of the Census for the data used in this research, and helpful discussion. We also wish to thank John Thompson, Henry Woltman and Jon Clark of the U.S. Bureau of the Census, and the referees, and the Associate Editor, for their many helpful comments. This research was supported in part by a grant from the National Science Foundation.

TECHNICAL APPENDIX A

A Review of the Hui and Walter Method

The Hui and Walter method was developed for the evaluation of diagnostic tests. The advantage of the technique is that, it allows the researcher to measure the error

rate in a given test, without requiring the comparison test to be error-free. To accomplish this task, the procedure uses two populations (or subpopulations) with different prevalences, to estimate the parameters. The data from such a study, can be summarized in a 2×2 table as given in Figure A below. This Table for a specific subpopulation, is indexed by the letter g . We will denote the frequency of cases from subpopulation g , that have a classification from the first test, of status i ($i = 1$ for those having the trait, and $i = 2$ for those not having the trait), and from the second test of status j ($j = 1$ or 2), by n_{ij} . Let π denote the true unknown prevalence rate of the trait, and let α_r and β_r denote the unknown false positive and false negative rates. These error rates are indexed by the letter r , where $r = 1$ corresponds to the outcome from the first test, and $r = 2$ for the second test, (which, in our context, $r = 1$ corresponds to the original interview, and $r = 2$ to a reinterview). The false positive rate, α_r refers to the probability, that the evaluation from the r -th test, will classify the person as positive when in truth the person should have been classified as negative. Similarly, the false negative rate, β_r , is the probability that evaluation from the r -th test will classify the case as negative, when the case has the trait. One (1) minus each of these parameters, reflects to the specificity and sensitivity of the test (or survey) classification procedures, respectively.

Test 1 Outcome (Original Survey)	Test 2 Outcome (Reinterview)		Total
	Positive	Negative	
Positive	Cell 1	Cell 3	$n_{1.}$
Negative	Cell 2	Cell 4	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

Figure A. Cross-classification of Test 1 and Test 2 Outcomes

Assuming the errors of the first and second tests are independent of each other (given the true state), the expected probabilities, denoted by P_{ij} associated with the cell frequencies given in Figure A, for a given subpopulation g are as follows:

For

$$\begin{aligned}
 \text{Cell 1 } P_{g11} &= \pi_g(1-\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(\alpha_{1,g}\alpha_{2,g}) \\
 \text{Cell 2 } P_{g21} &= \pi_g(\beta_{1,g})(1-\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(\alpha_{2,g}) \\
 \text{Cell 3 } P_{g12} &= \pi_g(1-\beta_{1,g})\beta_{2,g} + (1-\pi_g)(\alpha_{1,g})(1-\alpha_{2,g}) \\
 \text{Cell 4 } P_{g22} &= \pi_g(\beta_{1,g}\beta_{2,g}) + (1-\pi_g)(1-\alpha_{1,g})(1-\alpha_{2,g}). \quad (\text{A.1})
 \end{aligned}$$

From (A.1), we observe that we have a total of five parameters, but only three independent cell entries (or degrees of freedom), from which to estimate them. Therefore, the number of parameters must be reduced.

To reduce the parameters, Hui and Walter first, assume that, the proportion of cases with the trait, differs by subpopulation, which implies that, $\pi_1 \neq \pi_2$. Secondly, they require that two subpopulations can be found, such that the error rates for each test are the same for both subpopulations. The error rates associated with the two tests are allowed to differ. For two subpopulations, this implies that in (A.1), $\beta_r = \beta_{r,1} = \beta_{r,2}$, and $\alpha_r = \alpha_{r,1} = \alpha_{r,2}$, with $\beta_1 \neq \beta_2$, and $\alpha_1 \neq \alpha_2$. Under these conditions, the number of parameters reduces to six, (two prevalence rates, one for each subpopulation, and two error rates each for test 1 and test 2). Given that the two 2×2 tables contain six degrees of freedom, estimation is possible. Notice that if $\pi_1 = \pi_2$, and the error rates were the same in both subpopulations, then the probabilities in (A.1) would be the same for both subpopulations, so we would really have one table, and estimation would not be possible. Weighted nonlinear least squares estimates under the Hui and Walter model, can be computed using the Gauss Newton algorithm from the SAS Nonlinear Regression (NLIN) procedure. With this approach, one can express the observed frequencies, n_{ij} , in terms of the total sample size, $n_{..}$, multiplied by the probabilities in expression (A.1). Hui and Walter also present the closed formed estimators given in (A.2), expressed in terms of the observed cell probabilities denoted by p_{gij} .

$$\hat{\alpha}_r = \frac{(p_{r1} \cdot p_{\bar{r}1} - p_{r1} p_{\bar{r}1} + p_{211} - p_{111} + D)}{2E_r}$$

$$\hat{\beta}_r = \frac{(p_{r2} p_{\bar{r}2} - p_{r2} \cdot p_{\bar{r}2} + p_{122} - p_{222} + D)}{2E_r} \quad (\text{A.2})$$

where,

$$\bar{r} = 2 \text{ if } r = 1, \bar{r} = 1 \text{ if } r = 2$$

$$p_{g \cdot j} = \sum_{i=1}^2 p_{gij}, p_{gi \cdot} = \sum_{j=1}^2 p_{gij};$$

$$\hat{\pi}_g = \frac{1}{2} + \frac{[p_{g1} \cdot (p_{11} - p_{21}) + p_{g2} \cdot (p_{11} - p_{21}) + p_{211} - p_{111}]}{2D}$$

where,

$$D = \pm [(p_{11} \cdot p_{21} - p_{11} p_{21} + p_{111} - p_{211})^2 - 4(p_{11} - p_{21})(p_{111} p_{21} - p_{211} p_{11})]^{\frac{1}{2}}$$

with,

$$E_1 = p_{21} - p_{11}, E_2 = p_{21} - p_{11}.$$

Note that two distinct points exist in the solution set, for either a positive or a negative value of D ; however, only one of the values will yield reasonable estimates. Variances for

the estimators, derived from the estimated asymptotic information matrix, are given in Hui and Walter's (1980) paper.

TECHNICAL APPENDIX B

Adjusting the Reported Unemployment Rates

To evaluate the implications of the estimated error rates, we needed an expression for estimating the actual prevalence rates (the four π parameters), in terms of the estimated error rates and the observed prevalence rates (or sample frequencies), from a given survey. In this section, we present the formula for these computations. With this expression, we can use the BLS reported unemployed and employed prevalence rates, as the observed values to estimate the adjusted BLS prevalence rates. Such an expression is given in (B.1).

Note that in expression (B.1), we have deleted the g -th subscript from the π parameters, so that the expression represents the prevalence rates among the general population, males and females combined. Note that, in this study, we have assumed that the estimated error rates are equal for males and females.

$$\begin{bmatrix} \hat{\pi}_{y1} \\ \hat{\pi}_{y2} \end{bmatrix} = \begin{bmatrix} 1 - \hat{\beta}_{121} - \hat{\beta}_{131} - \hat{\beta}_{113} & \hat{\beta}_{112} - \hat{\beta}_{113} \\ \hat{\beta}_{121} - \hat{\beta}_{123} & 1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} \end{bmatrix}^{-1} \begin{bmatrix} \frac{n_{y1.}}{n_{y..}} - \hat{\beta}_{113} \\ \frac{n_{y2.}}{n_{y..}} - \hat{\beta}_{123} \end{bmatrix}. \quad (\text{B.1})$$

In this paper, we have three sets of observed values. We have two observed prevalence rates from the reinterview sample (which is a sub-sample of the full CPS sample), including the unreconciled reinterview sample data, and the reconciled reinterview data, from the response-bias study sample, and BLS reported prevalence rates, as observed from the full CPS original survey. We will concentrate our efforts on the first and last of these three sets of statistics, the unreconciled reinterview sample data, and the published BLS estimates. To keep these two sets separate, we will define,

$$U_y^R = \frac{n_{y1.}}{n_{y..}}$$

$$E_y^R = \frac{n_{y2.}}{n_{y..}} \quad (\text{B.2})$$

as the observed unemployed and employed prevalence rates, obtained from the CPS unreconciled reinterview sample data. The corresponding BLS reported prevalence rates based on the full CPS original survey weighted data, are defined by U_y^{BLS} and E_y^{BLS} .

Similarly, the observed unemployment rate among those in the labour force, from the unreconciled reinterview sample data, is denoted by UE_y^R , equal to U_y^R divided by $(U_y^R + E_y^R)$, and the observed BLS reported unemployment rate, is defined as UE_y^{BLS} .

Simplifying expression (B.1) in terms of the observed reinterview prevalence rates, U_y^R and E_y^R we find:

$$\begin{aligned}\hat{\pi}_{y1} &= \left\{ U_y^R - \hat{\beta}_{113} - \hat{\beta}_{112} U_y^R + \hat{\beta}_{113} \hat{\beta}_{112} - \hat{\beta}_{113} \hat{\beta}_{132} \right. \\ &\quad \left. - \hat{\beta}_{123} U_y^R - \hat{\beta}_{112} E_y^R + \hat{\beta}_{123} \hat{\beta}_{112} + \hat{\beta}_{113} E_y^R \right\} \\ &\quad \left\{ 1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} - \hat{\beta}_{121} (1 + \hat{\beta}_{132} + \hat{\beta}_{123} + \hat{\beta}_{113}) \right. \\ &\quad \left. - \hat{\beta}_{131} (\hat{\beta}_{112} + \hat{\beta}_{132} - 1) - \hat{\beta}_{113} (\hat{\beta}_{112} + \hat{\beta}_{132} - 1) + \hat{\beta}_{123} \hat{\beta}_{112} \right\} \\ \hat{\pi}_{y2} &= \left\{ -\hat{\beta}_{121} U_y^R + \hat{\beta}_{121} \hat{\beta}_{113} + \hat{\beta}_{123} U_y^R + E_y^R - \hat{\beta}_{123} - \hat{\beta}_{121} E_y^R \right. \\ &\quad \left. + \hat{\beta}_{122} \hat{\beta}_{123} - \hat{\beta}_{131} E_y^R + \hat{\beta}_{131} \hat{\beta}_{123} - \hat{\beta}_{123} E_y^R \right\} \\ &\quad \left\{ 1 - \hat{\beta}_{112} - \hat{\beta}_{132} - \hat{\beta}_{123} - \hat{\beta}_{121} (1 + \hat{\beta}_{132} + \hat{\beta}_{123} + \hat{\beta}_{113}) \right. \\ &\quad \left. - \hat{\beta}_{131} (\hat{\beta}_{112} + \hat{\beta}_{132} - 1) - \hat{\beta}_{113} (\hat{\beta}_{112} + \hat{\beta}_{132} - 1) + \hat{\beta}_{123} \hat{\beta}_{112} \right\}. \quad (\text{B.3})\end{aligned}$$

Using expression (B.3), we can compute estimates of the adjusted unemployment rate among those in the labour force from the reinterview survey, denoted by AUE_y^R equal to $\hat{\pi}_{yg1}$ divided by $(\hat{\pi}_{yg1} + \hat{\pi}_{yg2})$. Note the AUE_y^R can be expressed as follows:

$$\begin{aligned}AUE_y^R &= \left\{ -U_y^R + E_y^R + \hat{\beta}_{112} (U_y^R - \hat{\beta}_{113} + E_y^R) \right. \\ &\quad \left. + \hat{\beta}_{132} (U_y^R - \hat{\beta}_{113}) + \hat{\beta}_{123} (U_y^R - \hat{\beta}_{112}) - \hat{\beta}_{113} E_y^R \right\} \\ &\quad \left\{ U_y^R + \hat{\beta}_{113} (1 + \hat{\beta}_{112} - \hat{\beta}_{121} - \hat{\beta}_{123}) + \hat{\beta}_{112} (U_y^R + E_y^R - \hat{\beta}_{113}) \right. \\ &\quad \left. + \hat{\beta}_{121} (U_y^R + E_y^R - \hat{\beta}_{123}) - E_y^R + \hat{\beta}_{123} + \hat{\beta}_{131} (U_y^R - \hat{\beta}_{123}) \right\}. \quad (\text{B.4})\end{aligned}$$

Finally, to obtain the adjusted estimate of the BLS unemployment rate, denoted by AUE_y^{BLS} , we substitute the values of U_y^{BLS} for U_y^R and E_y^{BLS} for E_y^R , into expression (B.4). Note that the estimated standard errors of the estimates for AUE_y^{BLS} , presented in section four, were computed using a Taylor series approximation method, (Wolter 1985). As a first step in this process, we assumed the variance in the published estimates of U_y^{BLS} and E_y^{BLS} were negligible. While this is not true, this assumption greatly simplifies the computation of the variances, and captures the majority of the total variation. This assumption is supported by the fact, that the size of the variance of these estimates, given the large full CPS yearly sample sizes is negligible in comparison to the sampling error associated

with error rate estimates, which are based on the small unreconciled reinterview sample sizes. In summary, once the substitution of U_y^{BLS} for U_y^R , and E_y^{BLS} for E_y^R into expression (B.4) is completed, we assume that U_y^{BLS} and E_y^{BLS} are fixed known values in this equation. Finally, the sampling variance associated with the difference between the adjusted value and the published value, which defines the bias in the original estimate, is computed from the sum of the variances. Hence, by assuming the published value is sampling variance-free, the sampling variability associated with the difference or bias, is simply equal to the sampling variability associated with the adjusted value.

TECHNICAL APPENDIX C

Estimating Standard Errors of the Adjusted Unemployment Rates

For a complex function of several estimated parameters, the estimates of the variances associated with this function, can be computed using a Taylor series approximation as discussed by Wolter (1985). Suppose that the population parameter of interest is $Y = G(\Theta)$. Where Θ represents a n dimensional vector of population parameters, $\Theta = \{\theta_1, \dots, \theta_n\}$. If G possesses continuous second derivatives, in an admissible range for Θ and $\hat{\Theta}$, then Wolter (1985) presents the relationship:

$$\hat{Y} - Y = A + R(\hat{\Theta}, \Theta)$$

where,

$$\begin{aligned}A &= \sum_{k=1}^n \frac{\partial G(\Theta)}{\partial \theta_k} (\hat{\theta}_k - \theta_k) \\ R(\hat{\Theta}, \Theta) &= \sum_{k=1}^n \sum_{i=1}^n (1/2!) \frac{\partial^2 G(\Lambda)}{\partial \theta_k \partial \theta_i} (\hat{\theta}_k - \theta_k) (\hat{\theta}_i - \theta_i) \\ \hat{\Theta} &\leq \Lambda \leq \Theta. \quad (\text{C.1})\end{aligned}$$

The remainder term is often regarded of little consequence, and is eliminated from the relationship. Given the first order approximation, Wolter (1985) presents,

$$\begin{aligned}\text{MSE}(\hat{Y}) &= E[G(\hat{\Theta}) - G(\Theta)]^2 \\ &= \text{Var}(A) \\ &= \sum_{k=1}^n \sum_{i=1}^n \frac{\partial G(\Theta)}{\partial \theta_k} \frac{\partial G(\Theta)}{\partial \theta_i} \text{Cov}(\hat{\theta}_k, \hat{\theta}_i) \\ &= d \Sigma_{\hat{\Theta}} d^T \quad (\text{C.2})\end{aligned}$$

where d is a row vector of dimension n with the elements,

$$d_k = \left[\frac{\partial G(\Theta)}{\partial \theta_k} \right]. \quad (C.3)$$

Wolter calls this estimator, the first order approximation to the mean square error (equal to the sampling variance + the bias of the estimator squared). Higher order approximations can be developed, by retaining additional terms in the expansion. For purposes of variance estimation, we substitute the estimated covariance matrix for \sum_{θ} , and evaluate d at the estimated values of θ . Specifically, in our problem, we wish to estimate the variance associated with the function of the estimates in expression (C.4), given below.

$$\begin{aligned} G(\Theta) = & G(\hat{\beta}_{121}, \hat{\beta}_{131}, \hat{\beta}_{112}, \hat{\beta}_{132}, \hat{\beta}_{113}, \hat{\beta}_{123}, U_y^{\text{BLS}}, E_y^{\text{BLS}}) = \\ & \left\{ -U_y^{\text{BLS}} + E_y^{\text{BLS}} + \hat{\beta}_{112}(U_y^{\text{BLS}} - \hat{\beta}_{113} + E_y^{\text{BLS}}) \right. \\ & \left. + \hat{\beta}_{132}(U_y^{\text{BLS}} - \hat{\beta}_{113}) + \hat{\beta}_{123}(U_y^{\text{BLS}} - \hat{\beta}_{112}) - \hat{\beta}_{113}E_y^{\text{BLS}} \right\} \\ & \left\{ U_y^{\text{BLS}} + \hat{\beta}_{113}(1 + \hat{\beta}_{112} - \hat{\beta}_{121} - \hat{\beta}_{123}) + \hat{\beta}_{112}(U_y^{\text{BLS}} + E_y^{\text{BLS}} - \hat{\beta}_{113}) \right. \\ & \left. + \hat{\beta}_{121}(U_y^{\text{BLS}} + E_y^{\text{BLS}} - \hat{\beta}_{123}) - E_y^{\text{BLS}} + \hat{\beta}_{123} + \hat{\beta}_{131}(U_y^{\text{BLS}} - \hat{\beta}_{123}) \right\} \quad (C.4) \end{aligned}$$

To create the estimates, we have assumed that the values of U_y^{BLS} and E_y^{BLS} are fixed (*i.e.*, have a negligible sampling variance). Taking the partial derivatives of equation (C.4) with respect to the six error rates, and evaluating these expressions at the estimated values of the error rates, yield a vector d which depends on the values of the error rate estimates and the published BLS unemployed and employed proportions for each year of the study. With our original model, that assumes the error rates are fixed across each year, this d vector for the period of study, only varies from year-to-year for the published values. For illustrative purposes the estimated vector d for 1989 using the BLS published unemployed and employed prevalence rates of .0347 and .6329 is equal to:

$$\hat{d} = \begin{bmatrix} \hat{\beta}_{121} & .07851 \\ \hat{\beta}_{131} & .07558 \\ \hat{\beta}_{112} & -1.2918 \\ \hat{\beta}_{132} & -.04813 \\ \hat{\beta}_{113} & -.64214 \\ \hat{\beta}_{123} & .03884 \end{bmatrix}.$$

The estimated covariance matrix from our SAS NLIN analysis, which, based on the original model that assumes the error rates are fixed by year, and as such, is the same for all years under study, is given below.

\sum	β_{121}	β_{131}	β_{112}	β_{132}	β_{113}	β_{123}
β_{121}	0.000358	-4.7E-05	-3.5E-07	-2.6E-08	-3.9E-07	2.9E-07
β_{131}	-4.7E-05	0.000214	-1.7E-07	-5.2E-07	-1.4E-06	-2.8E-07
β_{112}	-3.5E-07	-1.7E-07	1.54E-06	2.14E-07	-2.3E-08	9.9E-10
β_{132}	-2.6E-08	-5.2E-07	2.14E-07	2.37E-06	-1.5E-08	-6.1E-08
β_{113}	-3.9E-07	-1.4E-06	-2.3E-08	-1.5E-08	2.4E-06	-8E-08
β_{123}	2.9E-07	-2.8E-07	9.9E-10	-6.1E-08	-8.0E-08	6.1E-06

Pre and post multiplying the vector d , by the estimated covariance matrix, yields an estimated variance for AUE^{BLS} for 1989 of 6.72 E-6 and a standard error of the estimate equal to .0026 (.26%) as given in Table 4.

REFERENCES

- ABOWD, J., and ZELLNER, A. (1985). Estimated gross labor force flows. *Journal of Economic and Business Statistics*, 3, 253-283.
- BAILAR, B.A. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.
- BIEMER, P.P., and FORSMAN, G. (1992). On the quality of reinterview data with application to the current population survey. *Journal of the American Statistical Association*, 87, 915-923.
- BUREAU OF LABOR STATISTICS (1993). Overhauling the population survey. *Monthly Labor Review*, 116, 9. Washington DC: U.S. Government Printing Office.
- BUREAU OF LABOR STATISTICS (1992). *Employment and Earnings*. 38, 8. Washington DC: U.S. Government Printing Office.
- CHUA, T.C., and FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.
- FORSMAN, G., and SCHREINER, I. (1991). The design and analysis of reinterview: an overview. In *Measurement Errors in Surveys*, (Eds. Paul Biemer *et al.*). New York: John Wiley and Sons.
- GASTWIRTH, J.L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDs antibodies test data. *Statistical Science*, 2, 213-238.
- HAUSMAN, J.A., and MORTON, S. (1994). Misclassification of a Dependent Variables in a Discrete Response Setting. Working paper, Department of Economics, MIT, Cambridge.
- HUI, S.L., and WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.
- HUGHES, J.J., and PERLMAN, R. (1984). *The Economics of Unemployment*. New York: Cambridge University Press.
- LEVY, P., and LEMESHOW, S. (1980). *Sampling for Health Professionals*. California: Lifetime Learning Publications.

- McKENNA, C.J. (1985). *Uncertainty and the Labor Market: Recent Developments in Job Search Theory*. New York: St. Martins Press.
- POTERBA, J.M., and SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.
- POTERBA, J.M., and SUMMERS, L.H. (1995). Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.
- RAO, J.N.K., and THOMAS, D.R. (1991). Chi-squared tests with complex survey data subject to misclassification error. In *Measurement Errors in Surveys*, (Eds. Paul Biemer *et al.*). New York: Wiley.
- SCHREINER, I., (1980). Reinterview Results From the CPS Independent Reconciliation Experiment (second quarter 1978 through third quarter 1979). Unpublished U.S. Bureau of the Census memorandum, May 7, 1980.
- SINCLAIR, M.D. (1994). Evaluating Reinterview Survey Methods for Measuring Response Errors. Phd. Dissertation, George Washington University, September.
- SINCLAIR M.D., and GASTWIRTH, J.L. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, (435) 961-969.
- SINGH, A.C., and RAO, J.N.K (1995). On the adjustment of gross flow estimate for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.
- U.S. BUREAU OF THE CENSUS (1963). *The Current Population Survey Reinterview Program: Some Notes and Discussion*. Technical Paper No. 6. Washington, D.C.: U.S. Government Printing Office.
- U.S. BUREAU OF THE CENSUS (1985). *Evaluating Censuses of Population and Housing*. Statistical Training Document #ISP-TR-5. Washington, D.C.: U.S. Government Printing Office, 1985.
- VACEK, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41, 959-968.
- WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

Use of Statistical Matching Techniques in Calibration Estimation

ROBBERT H. RENSSSEN¹

ABSTRACT

This article deals with an attempt to cross-tabulate two categorical variables, which were separately collected from two large independent samples, and jointly collected from one small sample. It was assumed that the large samples have a large set of common variables. The proposed estimation technique can be considered a mix between calibration techniques and statistical matching. Through calibration techniques, it is possible to incorporate the complex designs of the samples in the estimation procedure, to fulfill some consistency requirements between estimates from various sources, and to obtain fairly unbiased estimates for the two-way table. Through the statistical matching techniques, it is possible to incorporate a relatively large set of common variables in the calibration estimation, by means of which the precision of the estimated two-way table can be improved. The estimation technique enables us to gain insight into the bias generally obtained, in estimating the two-way table, by sole use of the large samples. It is shown how the estimation technique can be useful to impute values of the one large sample (donor source) into the other large sample (host source). Although the technique is principally developed for categorical variables Y and Z , with a minor modification, it is also applicable for continuous variables Y and Z .

KEY WORDS: Consistency between estimates; General regression estimator; Imputation; Multivariate auxiliary information; Two-way table.

1. INTRODUCTION

Most statistical surveys are conducted to obtain estimates of simple descriptive finite population parameters. The estimates are often presented in tabular form, with cells containing estimates of population totals or subgroup totals. Often, data are collected on an extensive set of variables, producing numerous results for these variables and their relationships. In order to save resources and decrease response burden, statistical bureaus wish to reduce sample sizes and shorten questionnaires. They resort to administrative data sources and existing large-scale sample surveys, or applying splitting questionnaire survey designs (see Raghunathan and Grizzle 1995). As a consequence, methods for combining distinct data sources have become a popular tool in the production of statistics. Combining data sources can be done in many different ways; two well-known techniques in survey sampling are statistical matching and calibration estimation.

Singh, Mantel, Kinack and Rowe (1993) describe statistical matching as a special case of imputation in which there are two distinct micro-data sources containing different information on different units. One data source serves as a host or recipient file to which new information is imputed for each record, using data from the other source, which is the donor file. More specifically, they consider a host file A, containing information on variables (X, Y) and a donor file B containing information on variables (X, Z) . The common variable X can be used to identify similar units in the two files. In general, statistical matching deals with the

problem of completing the records in file A, by imputing values for Z using the information on the (X, Z) relationship in file B. These imputed Z -values suffer from a serious limitation in that, the real relationship between Y and Z may be completely lost in the enriched host file. This limitation amounts to the so-called assumption of conditional independence between Y and Z given X . In order to get rid of this conditional independence assumption, Singh *et al.* (1993) consider a third data set (file C) representing auxiliary information about the full set (X, Y, Z) . For example, this data set could come from a small-scale specially conducted survey. They discuss several imputation methods to complete file A, by adding Z from file B using information from A, B, and C, on the joint relationships of X , Y , and Z . Singh *et al.* (1993) give many relevant references on statistical matching techniques. We only mention Rodgers (1984), Rubin (1986) and Paass (1986).

In Deville and Särndal (1992), calibration estimation is derived as a general technique to weight sample surveys, taking into account the complex design of the sample and auxiliary information obtained from external sources (see also Deville, Särndal, and Sautory 1993). The use of auxiliary information, *i.e.*, control variables, primarily aim at three goals: namely, reducing sampling variance, reducing bias due to non-response, and ensuring consistency between estimates from various sources with respect to the used control variables. There is an extensive body of literature on weighting methods in sample surveys. We refer to Bethlehem and Keller (1987), Alexander (1987), Lemaître and Dufour (1987), and Zieschang (1990).

¹ Robbert H. Renssen, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, Netherlands.

This article deals with the specific problem of how to estimate the cross-product between Y and Z (e.g., the two-way table between Y and Z in case these variables are categorical or the covariance between Y and Z in case these variables are continuous), using statistical matching techniques as well as calibration estimation. We assume that two data files A and B represent two large-scale sample surveys, possibly both obtained by a complex design. In order to weight the specially conducted small sample (file C), auxiliary information is derived from these large samples. It might be difficult to judge whether the large samples should be considered as suppliers of auxiliary information for the small sample, or vice versa. Through the statistical matching, it is possible to incorporate a large set of X -variables in the estimation procedure, despite the sample size of the small sample. The use of calibration estimation makes it possible to take account of the complex design of all samples in the estimation procedure, and to fulfill some consistency requirements. Most of the article is devoted to categorical Y and Z , because of the specific properties of these variables. For example, it is shown that the marginal counts of the estimated YZ -table, always coincide with estimates for the population totals of Y and Z , when the ordinary calibration estimator is applied with the X -variables as control variables, on the first and second large sample respectively. Nevertheless, the proposed method is also applicable for continuous Y and Z . Throughout this article it will be assumed that X may consist of several variables, which may be categorical and/or continuous. It is argued that when the X -variables are highly correlated with either Y or Z , then our estimation method gives relatively precise estimates for the cross-product between Y and Z , e.g., for the complete YZ -table when Y and Z are categorical.

The proposed estimation procedure closely resembles a method presented in Singh *et al.* (1993, Section 2) to estimate a correlation coefficient between Y and Z . These variables are assumed to be univariate in this article. Our method, however, differs from theirs in that it incorporates the complex designs of all data sources in the estimation procedure and that it uses the large data sources more efficiently in estimating population parameters from the small data source. When Y and Z are categorical, and there is no linear correlation between X and Y as well as between X and Z , then our method corresponds to incomplete post-stratification (Deville and Särndal 1992, Bethlehem and Keller 1987). On the other hand, if Y is perfectly correlated with X , then our method gives an estimated two-way table between Y and Z which corresponds to an estimated two-way table that would have been obtained from file B if first the Y -values were imputed. A similar result holds if Z and X are perfectly correlated.

Although combining distinct data sources across common variables may be fruitful from a theoretical point of view, in practice, complications may arise because common variables in the strict sense are not easily found,

mainly due to discrepancies between definitions, methods of observation, and reference period. These complications may be reduced if the survey processes involved, are harmonized at an early stage. A promising application of the use of common variables, lies in integrated survey designs, such as the Dutch Household Survey on Living Conditions, see van Tuinen (1995), Bakker and Winkels (1998), Winkels and Everaers (1998), and Hofmans (1998). The questionnaire design of this survey has a three-shell structure. The first shell contains questions on demographic and socioeconomic issues, and level of education. The second shell contains a few easy to answer core questions, on every relevant aspect of living conditions. The questions in the third shell also concern living conditions, but they are more exhaustive than the questions in the second shell. In order to shorten the time it takes to answer, the third shell questionnaire is split. Each respondent has to fill in the complete questionnaire of the first and second shell and one sub-questionnaire of the third shell. On account of the third shell, the sample is split into sub-samples associated with each sub-questionnaire. The sampling design of each sub-sample can be described as two-phase sampling for the general regression estimator.

The organization of this article is as follows. The theoretical framework is developed in Section 2. For this purpose it is convenient to discuss a calibration estimator for the small sample, obtaining auxiliary information from two distinct registrations instead of two distinct large samples. One registration contains values on X and Y and the other registration on X and Z . Sections 2.1 to 2.4 deal with categorical Y - and Z -variables. In Section 2.1, the registrations are used to obtain a first synthetic estimate of the YZ -table by regression methods of imputation. It is shown that this synthetic two-way table has some interesting properties. In Section 2.2 we propose a set of calibration equations to weight the small sample, based on these properties. We briefly discuss its relationship to complete and incomplete post-stratification. A numerical illustration is given in Section 2.3. The linkage to statistical matching techniques as discussed in Singh *et al.* (1993) is given in Section 2.4. The treatment of categorical Y and Z is unnecessary and restrictive. In Section 2.5, it is shown that the proposed weighting technique is also applicable for continuous Y and Z or for continuous Y and categorical Z . In Section 3, the technique is modified, using auxiliary information from two distinct large samples instead of two registrations. By means of a simulation study, the modified weighting method is compared to the traditional incomplete two-way stratification. Finally, Section 4 contains some concluding remarks.

2. COMBINING REGISTRATIONS ACROSS COMMON VARIABLES

Consider a finite population $\Omega = \{1, \dots, N\}$ of N persons and suppose there are two registrations available of these

persons. The first registration contains of each person k , a record with scores y_k and x_k of the variables Y and X respectively, and the second registration of each person k , a record with scores z_k and x_k of the variables Z and X respectively, $k = 1, \dots, N$. Obviously, the variable X is present in both registrations. We note that the records from both registrations correspond to the same finite population. The process of merging these registrations, would be like exact matching if X is used to compare the records in the one registration with those in the other registration, in an effort to determine which pairs of records relate to the same population unit (see Fellegi and Sunter 1969). In this article we will proceed differently.

2.1 Formulating the Synthetic Population Totals

Let Y denote education with p categories and Z denote employment with q categories. Then y_k is a vector of order p , representing p dummy variables. Each dummy variable corresponds to a specific category; it equals 1 if person k belongs to that category, otherwise it equals 0. Analogously defined, z_k is a vector of order q . Further, X may be the result of a complete or incomplete crossing (stratification) of a number of characteristics (*e.g.*, sex, age, region, marital status, *etc.*). The scores x_k are vector valued, of order r . In case X consists of a complete stratification, x_k represents r dummy variables. In the remaining of this article, r should be considered large in comparison with $p \times q$. The population totals for Y and Z are the marginal frequency distributions with respect to education and employment. Using the common variable X , predictions for Y and Z can be defined with a multiple linear regression model:

$$\hat{y}_k = B'x_k, \quad k = 1, \dots, N,$$

and

$$\hat{z}_k = A'x_k, \quad k = 1, \dots, N,$$

where B and A are the ordinary least squares regression coefficients satisfying the normal equations

$$\left(\sum_{k=1}^N x_k x_k' \right) B = \sum_{k=1}^N x_k y_k' \quad (1)$$

and

$$\left(\sum_{k=1}^N x_k x_k' \right) A = \sum_{k=1}^N x_k z_k'. \quad (2)$$

The superscript ' t ' denotes transposition. This model is called a linear probability model, (see Maddala 1983, chap. 2). There are more elegant models, such as probit and logit models, to predict binary variables. However, we are not interested in the predictions themselves, but in the

synthetic population totals of these predictions. These totals appear to have nice properties if the linear prediction model is used, and for this reason the model can be justified. Note that B is calculated from the first registration and A from the second one. By means of the common variable X and the regression coefficients B and A , we construct a synthetic registration, which contains a record of each person k with scores x_k , $B'x_k$, and $A'x_k$. In fact, either y_k or z_k may be added to this registration, but for our purposes this addition appears to be superfluous (see next paragraph). If there exists a vector a of order r of fixed numbers such that $a'x_k = 1$ for all k , then the population totals of the new variables $B'x_k$ and $A'x_k$ equal the population totals of the corresponding original variables (see *e.g.*, Bethlehem and Keller 1987). This can be shown easily by first pre-multiplying the normal equations (1) and (2) by a' and subsequently substituting $a'x_k = 1$ into the resulting equations.

From the synthetic registration, a synthetic two-way table can be defined by $\sum_{k=1}^N (B'x_k)(A'x_k)'$. This synthetic two-way table can be considered as an approximation of the (simultaneous) frequency distribution $\sum_{k=1}^N y_k z_k'$. Using the normal equations (1) and (2), the following identities can be derived:

$$\begin{aligned} \sum_{k=1}^N (B'x_k)(A'x_k)' &= \sum_{k=1}^N y_k (A'x_k)' \\ &= \sum_{k=1}^N (B'x_k)z_k'. \end{aligned}$$

Clearly, the crossings between $B'x_k$ and $A'x_k$, y_k and $A'x_k$, or $B'x_k$ and z_k , all result in identical synthetic two-way tables. Therefore, it suffices to consider only $\sum_{k=1}^N (B'x_k)(A'x_k)'$, and delete either y_k or z_k in the synthetic registration. The difference between the real frequency distribution between Y and Z and its synthetic "approximation", can be obtained from the following decomposition

$$\begin{aligned} \sum_{k=1}^N y_k z_k' &= \sum_{k=1}^N (B'x_k)(A'x_k)' + \\ &\quad \sum_{k=1}^N (y_k - B'x_k)(z_k - A'x_k)'. \end{aligned} \quad (3)$$

Note the strong resemblance with the ordinary variance decomposition in regression analysis (see *e.g.*, Searle 1971). If either $B'x_k = y_k$ or $A'x_k = z_k$ for all k , then the two-way table derived from the synthetic registration, equals the real simultaneous frequency distribution between Y and Z .

Let l be a vector of appropriate order consisting of ones, and note that $l'y_k = 1$ and $l'z_k = 1$ for all k . If there exists a constant a such that $a'x_k = 1$ for all k , then we also have

$$l' \hat{y}_k = l' B' x_k = l' \left(\sum_{k=1}^N y_k x_k' \right) \left(\sum_{k=1}^N x_k x_k' \right)^{-1} x_k =$$

$$a' \left(\sum_{k=1}^N x_k x_k' \right) \left(\sum_{k=1}^N x_k x_k' \right)^{-1} x_k = a' x_k = 1$$

for all k , and similarly $l' \hat{z}_k = l' A' x_k = 1$ for all k . It follows that

$$l' \sum_{k=1}^N (B' x_k) (A' x_k)' l = \sum_{k=1}^N (A' x_k)' l = \sum_{k=1}^N z_k' \quad (4)$$

and

$$\sum_{k=1}^N (B' x_k) (A' x_k)' l = \sum_{k=1}^N (B' x_k) = \sum_{k=1}^N y_k. \quad (5)$$

So, the row and column totals of the synthetic two-way table, equal the corresponding marginal population counts with respect to Y and Z .

What remains to consider, is the condition $a' x_k = 1$ for all k , for some constant a . This condition is satisfied if X represents a categorical variable. More generally, the condition is always satisfied if the vector X can be partitioned into two sub-vectors, one of which represents a categorical variable.

2.2 Formulating the Constraints in Calibration Estimation

Suppose a probability sample s of size n is drawn from the finite population $\Omega = \{1, \dots, N\}$ according to a sampling design $p(s)$ such that the first and second order inclusion probabilities $\Pr(k \in s) = \pi_k$ and $\Pr(k, l \in s) = \pi_{kl}$ are strictly positive. For each $k \in s$ the vector of scores (x_k, y_k, z_k) is observed. Two distinct registrations are available to provide auxiliary information. The first registration contains for each $k \in \Omega$, records with scores on x_k and y_k , the second registration contains for each $k \in \Omega$, scores on x_k and z_k . The objective is to estimate the YZ -table from the sample s , using auxiliary information from both registrations. There exists a wide range of weighting type estimators in the presence of multivariate auxiliary information. In Särndal, Swensson and Wretman (1992), the general regression estimator is extensively discussed. It implicitly defines sample weights, which reproduce the known population totals of the auxiliary variables, used as control variables in the estimator. Such a consistency property is attractive if the auxiliary information is used both for publication and for weighting. As a generalization of the general regression estimator, the calibration estimator is developed (Deville and Särndal 1992 and Deville *et al.* 1993).

To be specific, let G be a real valued function as defined in Deville *et al.* (1993) and consider the following weighting type estimator for our YZ -table:

$$\hat{T} = \sum_{k=1}^n w_k (y_k z_k'), \quad (6)$$

where w_k is a scalar, representing a weight assigned to person $k \in s$. Denote $d_k = \pi_k^{-1}$. A calibration estimator for the YZ -table uses weights which are obtained by minimizing $\sum_{k=1}^n d_k G(w_k/d_k)$ with respect to w_k subject to a set of constraints on w_k for any particular sample s . We first consider the following set of constraints:

$$\sum_{k=1}^n w_k y_k = \sum_{k=1}^N y_k \quad \text{and} \quad \sum_{k=1}^n w_k z_k = \sum_{k=1}^N z_k. \quad (I)$$

This (first) set of constraints only uses the (marginal) counts with respect to Y and Z . No use is made of the common variable X . One of the $p + q$ equations is redundant, so to solve the minimization problem, one equation can be deleted. For $G(w_k/d_k) = (w_k/d_k - 1)^2$, the resulting calibration estimator corresponds to incomplete two-way stratification as defined in Bethlehem and Keller (1987). By taking $G(w_k/d_k) = 1 + w_k/d_k (\log(w_k/d_k) - 1)$, the classical raking ratio estimator is obtained (see *e.g.*, Oh and Scheuren 1987). Copeland, Peitzmeier and Hoy (1987) have compared these methods, based on data of the Current Population Survey. They conclude that the estimates produced by the two methods are very similar. In Deville *et al.* (1993), two other distance functions are discussed, which are especially interesting in view of the problem of extreme weights. Estimating two-way tables with constraints on the marginal counts, is frequently performed in sample surveys. Often, the constraints on the marginal counts are required for two reasons. The first reason is to reduce sampling error and sampling bias, and the second reason is to meet consistency requirements with published population counts.

Suppose that x_k is categorical with r categories. Since population information about the crossings between Y and X , and the crossings between Z and X are available, we may also consider the following set of constraints:

$$\sum_{k=1}^n w_k (y_k x_k') = \sum_{k=1}^N y_k x_k' \quad \text{and}$$

$$\sum_{k=1}^n w_k (z_k x_k') = \sum_{k=1}^N z_k x_k'.$$

The number of non-redundant constraints in this set equals $r(p + q - 1)$. For large r , this set may be not feasible because it contains too many constraints in comparison with

the sample size. Only if r is small, the set may be of practical interest. In the remaining of this article, this set of constraints will be disregarded.

In view of incorporating a large set of common variables in the weighting procedure, we consider a set of constraints, which exploits the bivariate population information that we have in the synthetic table:

$$\sum_{k=1}^n w_k (B^t x_k) (A^t x_k)^t = \sum_{k=1}^N (B^t x_k) (A^t x_k)^t. \quad (\text{II})$$

This (second) set of constraints is a straightforward application of the theory of calibration estimators. Population totals of the crossing between $B^t x_k$ and $A^t x_k$ are known, so these crossings are taken as auxiliary variables to formulate the set of constraints. Evidently, for large r , the number of non-redundant constraints remains bounded by $p \times q$. A major disadvantage of the resulting calibration weights is that, they do not necessarily reproduce the (marginal) population counts with respect to Y and Z , when applying these weights to y_k and z_k respectively. In other words, the resulting calibration weights do not necessarily satisfy the first set of constraints. Especially, if this set of constraints is formulated in view of consistency requirements, this is a serious drawback.

Therefore, as an alternative, we consider a third set of constraints:

$$\sum_{k=1}^n w_k (y_k z_k^t - (y_k - B^t x_k)(z_k - A^t x_k)^t) = \sum_{k=1}^N (B^t x_k) (A^t x_k)^t \quad (\text{III})$$

Assuming that there exists a constant a , such that $a^t x_k = 1$ for all k , this set of constraints meets the consistency objective. Let l denote a vector of ones of appropriate order and recall that $l^t y_k = l^t B^t x_k = l^t z_k = l^t A^t x_k = 1$ for all k , $B^t \sum_{k=1}^N x_k = \sum_{k=1}^N y_k$, and $A^t \sum_{k=1}^N x_k = \sum_{k=1}^N z_k$. By pre-multiplying the third set of equations on both sides with l^t , we obtain the first set of constraints with respect to Z , and post-multiplying the third set on both sides with l gives the first set of constraints with respect to Y . The resulting calibration estimator can be expressed as

$$\hat{T} = \sum_{k=1}^n w_k (y_k z_k^t) = \sum_{k=1}^N (B^t x_k) (A^t x_k)^t + \sum_{k=1}^N w_k (y_k - B^t x_k)(z_k - A^t x_k)^t.$$

Clearly, this estimator obeys the decomposition given by (3). It equals the synthetically defined two-way table plus an adjustment term. This adjustment term is a calibration estimate for the difference between the real frequency

distribution between Y and Z and the synthetically defined two-way table. Similarly to the second set of constraints, the number of non-redundant constraints in the third set is bounded by $p \times q$.

An important special case is $G(w_k/d_k) = (w_k/d_k - 1)^2$. Then each estimated cell is a general regression estimate with $(y_k z_k)$, $\text{vec}(B^t x_k x_k^t A)$, and $\text{vec}(y_k z_k^t - (y_k - B^t x_k)(z_k - A^t x_k)^t)$ as control variables in case of the first, second, and third set of constraints respectively. Analytical formulas for the design variance of the general regression estimator, are given in *e.g.*, Särndal *et al.* (1992, chap. 6). In fact, these formulas are approximations for large sample sizes. In Deville and Särndal (1992), sufficient conditions are given under which these approximations are valid for calibration estimators in general.

In Deville *et al.* (1993), complete post-stratification is described as a calibration method for which all population counts with respect to the cross-classifications, are used in the set of constraints. An elaboration of complete post-stratification, results in the ordinary post-stratification estimator, regardless of the distance function G . As an alternative, incomplete post-stratification is described as a calibration method, in which less detailed than a complete knowledge of all cell counts, is used in the constraint set. The calibration estimator defined under the first set of constraints, is a commonly used example of incomplete post-stratification. Several cases are discussed, in which incomplete post-stratification is preferable to complete post-stratification. Two of them are, lack of population information and, some zero or extremely small cell counts (see also Oh and Scheuren 1987). The calibration estimator defined under the second and third set of constraints, corresponds to complete post-stratification in the sense that, all crossings are used as auxiliary information. Except when a perfect linear relationship exists either between Y and X , or between Z and X , the method differs from complete post-stratification in using synthetic population totals instead of real population counts. Complete post-stratification gives unstable results, if some sample cells have only few observations. In such situations, incomplete post-stratification is of practical interest. Similarly, the calibration estimator under the second and third set of constraints may be unstable. Analogously to incomplete post-stratification, one might consider using an incomplete crossing in the constraints instead.

2.3 A Numerical Illustration

We illustrate the calibration estimator under the three different sets of constraints by means of a hypothetical example. The example is based on real data from a sample on behalf of the Dutch National Travel Survey (1994). The sampling design is roughly a self-weighted cluster sample of addresses. All persons living in a selected address, are included in the sample. The net sample size is approximately 80,000 persons within 34,000 addresses. From this sample, two hypothetical registrations of approximately

$N = 80,000$ persons are constructed. In the one registration, age is registered (in six categories), and in the other registration, car ownership (in two categories). The common variable between the registrations is a key number for addresses, resulting in $r = 34,000$ categories for the X -variable. For this particular example the synthetic two-way table simplifies to

$$\sum_{k=1}^N (B^t x_k)(A^t x_k)^t = \sum_{j=1}^r N_j \bar{y}_j \bar{z}_j^t,$$

where N_j denotes the size of the j -th address, \bar{y}_j the mean of the six age categories of the j -th address, and \bar{z}_j the mean of the two car ownership categories of the j -th address.

In order to calculate the synthetic two-way table, both registrations are combined as follows. Firstly, they are sorted according to the key number for addresses. Secondly, the address counts of the six age categories and the two car ownership categories are calculated. Thirdly, each address count of age, is linked with its corresponding address count of car ownership. By means of this synthetic registration of $r = 34,000$ addresses, the synthetic two-way table can be calculated. The result is shown in Table 1. This table can be considered as a first approximation of the real frequency distribution between age and car ownership. A sufficient condition for a close approximation, is homogeneity with respect to either age or car ownership within all addresses, *i.e.*, all persons at the same address should either be in the same age category or in the same car ownership category. For most (multiple) person addresses, this seems to be an unlikely proposition. It follows from equations (4) and (5) that the row and column totals in table 1 coincide with the real (marginal) population counts of age and car ownership respectively.

By means of a simple random sample of $n = 1000$ persons, the population cell counts are estimated using a general regression estimator. Three sets of auxiliary variables are used, in accordance with the three sets of constraints mentioned in the previous section. The estimated tables are given below (for convenience we have taken the quadratic distance measure: $G(w_k/d_k) = (w_k/d_k - 1)^2$). The corresponding estimated standard deviations are within parenthesis. These estimates are based on the usual variance formulas of the general regression estimator, see Särndal *et al.* (1992, chap. 6).

Table 1
Synthetic Population Totals for Crossings Between Age
and Car Ownership

	1	2	3	4	5	6	total
yes	3461	1659	5739	10770	6536	3334	31499
no	9827	4692	7902	17102	6424	5389	51336
total	13288	6351	13641	27872	12960	8723	82835

Table 2

Estimated Population Totals for Crossings Between Age and
Car Ownership, Satisfying the First Set of Constraints

	1	2	3	4	5	6	total
yes	0 ₍₀₎	0 ₍₀₎	4968 ₍₄₂₃₎	15414 ₍₅₄₃₎	7518 ₍₄₅₈₎	3599 ₍₃₇₅₎	31499
no	13288 ₍₀₎	6351 ₍₀₎	8673 ₍₄₂₃₎	12458 ₍₅₄₃₎	5422 ₍₄₅₈₎	5124 ₍₃₇₅₎	51336
total	13288	6351	13641	27872	12960	8723	82835

Table 3

Estimated Population Totals for Crossings Between Age and
Car Ownership, Satisfying the Second Set of Constraints

	1	2	3	4	5	6	total
yes	0 ₍₀₎	0 ₍₀₎	4791 ₍₄₃₅₎	13826 ₍₈₁₁₎	6887 ₍₄₉₄₎	3421 ₍₃₂₁₎	28923 ₍₁₀₀₅₎
no	14385 ₍₇₈₂₎	7012 ₍₅₉₅₎	8118 ₍₅₆₃₎	12893 ₍₇₉₆₎	5853 ₍₄₆₄₎	5654 ₍₃₀₆₎	53912 ₍₁₀₀₅₎
total	14385 ₍₇₈₂₎	7012 ₍₅₉₅₎	12908 ₍₆₀₃₎	26718 ₍₉₅₈₎	12739 ₍₄₁₉₎	9074 ₍₁₇₇₎	82835

Table 4

Estimated Population Totals for Crossings Between Age and
Car Ownership, Satisfying the Third Set of Constraints

	1	2	3	4	5	6	total
yes	0 ₍₀₎	0 ₍₀₎	5501 ₍₂₂₆₎	15647 ₍₂₂₇₎	6898 ₍₁₇₇₎	3453 ₍₇₈₎	31499
no	13288 ₍₀₎	6351 ₍₀₎	8139 ₍₂₂₆₎	12224 ₍₂₂₇₎	6062 ₍₁₇₇₎	5270 ₍₇₈₎	51336
total	13288	6351	13641	27872	12960	8723	82835

In Table 2 the population counts are estimated according to the ordinary incomplete two-way stratification (Bethlehem and Keller 1987). There are no young people (age category 1 and 2) owning a car, observed in the sample, which is likely to be representative for the population, so these cells are estimated by zero. Due to the consistency requirements, *i.e.*, the first set of constraints, it follows that the estimated cell counts of young people without a car equal the corresponding marginal cell counts. An attempt to improve Table 2, is to use the common variable address in the weighting procedure. In Table 3, the cell estimates are given according to the second set of constraints. As already mentioned in the previous section, the estimated row and column totals may differ from the real population counts. A comparison between Table 2 and Table 3 shows that these differences can be considerable. In addition, almost all estimated cell counts in Table 2 have smaller estimated standard deviations than the corresponding estimated cell counts in Table 3. So, the second set of constraints gives quite unsatisfactory results. The third set of constraints covers the first set of constraints. This implies 1) consistency of the estimated marginal cell counts with respect to the corresponding known population cell counts, and 2) smaller asymptotic variances of all estimated cell counts. The results are shown in Table 4. Indeed, the estimated marginal cell counts are consistent, and the estimated standard deviations are at most half of the corresponding standard estimates given in Table 2.

2.4 Imputing Values of the one Registration into the Other Registration

Until now, we have developed a weighting method to estimate a two-way table between two variables, which are registered in two distinct registrations. Often, one is interested not only in estimated two-way tables, or more generally, estimated linear relations, but in complete registrations in which both variables are simultaneously registered. Users of statistics find such complete data-bases easy to analyze. The creation of such enriched registrations can be seen as a special case of imputation. One registration serves as a host or recipient source, and the other as a donor source. Assuming the second registration to be the donor source, the problem is imputing Z -values from the second registration, into the first registration using the estimated two-way table discussed in Section 2.2, as auxiliary information. Statistical matching problems using data from a third data source, have already been considered by Rubin (1986) and Paass (1986). Singh *et al.* (1993) gives a review of their methods. In addition, they propose some modifications to Rubin's (1986) and Paass's (1986) methods. Our imputation method is based on the regression method suggested by Rubin (1986) and Singh *et al.* (1993).

After having defined predictors for the Z -variables by means of the regression model

$$\hat{z}_k = A'x_k, \quad k = 1, \dots, N,$$

where A is given by (2), we define new predictions for these variables by means of the enlarged regression model

$$\tilde{z}_k = \alpha_1'x_k + \alpha_2'y_k, \quad k = 1, \dots, N,$$

with

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \left[\sum_{k=1}^N \begin{pmatrix} x_k x_k' & x_k y_k' \\ y_k x_k' & y_k y_k' \end{pmatrix}^{-1} \right] \left[\sum_{k=1}^N \begin{pmatrix} x_k z_k' \\ y_k z_k' \end{pmatrix} \right].$$

Using well-known results about partial regression coefficients in the general linear model (see *e.g.*, Seber 1977), α_1 and α_2 can be expressed as

$$\alpha_1 = A - B\alpha_2$$

and

$$\alpha_2 = \left[\sum_{k=1}^N (y_k - B'x_k)(y_k - B'x_k)' \right]^{-1} \times \left[\sum_{k=1}^N (y_k - B'x_k)(z_k - A'x_k)' \right],$$

where B and A are given by (1) and (2) respectively. They can be calculated from the first and second registration. The partial regression coefficients should be estimated from the third source. We suggest

$$\hat{\alpha}_1 = A - B\hat{\alpha}_2$$

and

$$\hat{\alpha}_2 = \left[\sum_{k=1}^N (y_k - B'x_k)(y_k - B'x_k)' \right]^{-1} \times \left[\sum_{k=1}^n w_k (y_k - B'x_k)(z_k - A'x_k)' \right],$$

where w_k are calibration weights which are discussed in Section 2.2. Based on these estimates we define new predictions for the Z -values:

$$\hat{\tilde{z}}_k = \hat{\alpha}_1'x_k + \hat{\alpha}_2'y_k = A'x_k + \hat{\alpha}_2'(y_k - B'x_k), \quad k = 1, \dots, N. \quad (7)$$

These new predictions equal the old predictions (see Section 2.1) plus an adjustment term. This adjustment term depends on the difference between the Y -value and its (old) prediction. It can be viewed as an attempt to improve the prediction for Z , however, and more important, it is a means to reconstruct the weighting type estimator under the third set of constraints (Section 2.2). Indeed, the following equality holds:

$$\sum_{k=1}^N y_k \hat{\tilde{z}}_k = \sum_{k=1}^N (B'x_k)(A'x_k)' + \sum_{k=1}^n w_k (y_k - B'x_k)(z_k - A'x_k)'.$$

This is just the weighting type estimator under the third set of constraints, if the corresponding calibration weights are used to estimate α_2 . It is easy to show that

$$\sum_{k=1}^N x_k \hat{\tilde{z}}_k = \sum_{k=1}^N x_k \hat{z}_k = \sum_{k=1}^N x_k z_k'.$$

So, also the XZ -table can be reconstructed. At the beginning of this section, we assumed the second registration to be the donor source. This choice was arbitrary. If the Y -values were imputed instead of the Z -values, we would have obtained an identical estimate for the YZ -table. In addition, the XY -table could have been reconstructed.

The new predictions for the Z -values can be used for imputation. Singh *et al.* (1993) give algorithms for imputation using regression models. These Z -values can be imputed in the first registration in two steps. In the first step, the predictions given by (7) are calculated for each

(x_k, y_k) in the first registration. We have shown that the crossings between the Y -values and these predicted Z -values, can be considered as weighting type estimators. However, the calculated predictions have in general no realistic values, and therefore the first step is followed by a second step. In the second step, each predicted Z -value in the first registration is replaced by a live Z -value from the second registration, which is nearest under some Euclidean distance in (X, Z) .

2.5 Estimating Cross-Products for Continuous Y - and Z -Variables

The consistency property of the third set of constraints (Section 2.2) also hold with respect to continuous Y - and Z -variables, provided that there exist constants a_y and a_z of proper order, such that $a_y' y_k = 1$ and $a_z' z_k = 1$ for all k . To see this, we slightly extend the results of Section 2.1. First note that

$$a_y' B' x_k = a_y' \sum_{k=1}^N y_k x_k' \left(\sum_{k=1}^N x_k x_k' \right)^{-1} x_k =$$

$$a' \sum_{k=1}^N x_k x_k' \left(\sum_{k=1}^N x_k x_k' \right)^{-1} x_k = a' x_k = 1$$

(it is still assumed that there exists a constant a such that $a' x_k = 1$ for all k). Similarly, it holds that $a_z' A' x_k = 1$. The equivalent equations of (4) and (5) for the continuous case are readily obtained. Consequently, pre-multiplying both sides of (III) with a_y' gives $\sum_{k=1}^N w_k z_k' = \sum_{k=1}^N z_k'$ and post-multiplying both sides of (III) with a_z yields $\sum_{k=1}^N w_k y_k = \sum_{k=1}^N y_k$. So, the third set of constraints meets the consistency objective, *i.e.*, the calibration equation of the first set of constraints, for quite general Y - and Z -variables. We will give two examples.

In the first example we take $y_k = (1, y_{2k})'$ and $z_k = (1, z_{2k})'$, where both y_{2k} and z_{2k} are assumed to be continuous. By taking $a_y = a_z = (1, 0)'$ we see that $a_y' y_k = a_z' z_k = 1$ for all k . The cross-product between Y and Z equals

$$\sum_{k=1}^N y_k z_k' = \begin{pmatrix} N & \sum_{k=1}^N z_{2k} \\ \sum_{k=1}^N y_{2k} & \sum_{k=1}^N y_{2k} z_{2k} \end{pmatrix},$$

from which the covariance between y_{2k} and z_{2k} is easily derived. This cross-product can be estimated using the third set of constraints. An elaboration of this set gives the following four constraints for this particular example:

$$\sum_{k=1}^n w_k = N, \sum_{k=1}^n w_k y_{2k} = \sum_{k=1}^N y_{2k}, \sum_{k=1}^n w_k z_{2k} = \sum_{k=1}^N z_{2k},$$

and

$$\sum_{k=1}^n w_k (y_{2k} z_{2k} - (y_{2k} - B_2' x_k) (z_{2k} - A_2' x_k)) =$$

$$\sum_{k=1}^N (B_2' x_k) (A_2' x_k),$$

where the regression coefficients are given by

$$B_2 = \left(\sum_{k=1}^N x_k x_k' \right)^{-1} \sum_{k=1}^N x_k y_{2k}$$

and

$$A_2 = \left(\sum_{k=1}^N x_k x_k' \right)^{-1} \sum_{k=1}^N x_k z_{2k}.$$

If one is specially interested in the correlation coefficient between y_{2k} and z_{2k} , then following constraints may be considered in addition:

$$\sum_{k=1}^n w_k y_{2k}^2 = \sum_{k=1}^N y_{2k}^2 \text{ and } \sum_{k=1}^n w_k z_{2k}^2 = \sum_{k=1}^N z_{2k}^2.$$

In the second example, we suppose that $y_k = (1, y_{2k})'$, where y_{2k} may be continuous, and z_k is categorical with q categories. By taking $a_y = (1, 0)'$ and $a_z = l$, where l is a vector of ones of proper order, we see that $a_y' y_k = a_z' z_k = 1$ for all k . The cross-product between Y and Z is

$$\sum_{k=1}^N y_k z_k' = \begin{pmatrix} N_1 & N_2 & \cdots & N_q \\ \sum_{k \in C_1} y_{2k} & \sum_{k \in C_2} y_{2k} & \cdots & \sum_{k \in C_q} y_{2k} \end{pmatrix},$$

where C_h denotes the set of population elements belonging to the h -th category of Z , and N_h the size of C_h . It is ensured that the calibration weights according to the third set of constraints, satisfy the 'marginal' calibration equations $\sum_{k=1}^n w_k z_k = \sum_{k=1}^N z_k = (N_1 \dots N_q)'$ and $\sum_{k=1}^n w_k y_{2k} = \sum_{k=1}^N y_{2k}$, which both may be of interest in view of consistency requirements.

3. COMBINING INDEPENDENT SAMPLES ACROSS COMMON VARIABLES

In the previous section, we have presented a method for combining two registrations across common variables, using auxiliary information from a small sample. In this section, the method is adjusted by combining two independent samples. We consider a complete registration of persons, two large-scale sample surveys, and a small-scale sample survey. The registration contains a limited set of variables such as sex, age, region, and marital status. These

variables are denoted by X . In the one large sample, the variables Y , U , and X are observed, and in the other large sample, the variables Z , U , and X . In the small sample all variables, *i.e.*, Y , Z , U , and X , are observed. The small sample could come from a specially conducted small-scale survey, or from sample overlap of the large-scale surveys. In Figure 1, the data sources are schematically given. For convenience, it is assumed that all samples correspond to different units, *i.e.*, it is assumed that there is no sample overlap.

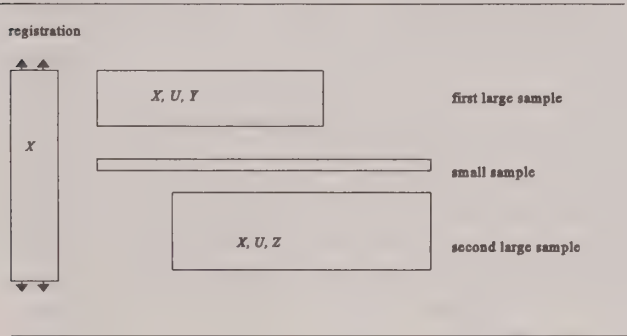


Figure 1. Overview of the Several Data Sources

The common variables X and U are partitioned into $C = (X \ U)$, where X denotes the set of common variables with known population totals, and U denotes the set of common variables with unknown population totals. All samples may be drawn by some complex sampling design. Both Y and Z are assumed to be categorical, however, as in Section 2.5, the suggested weighting methods are also applicable for continuous Y and Z . The purpose is to estimate the two-way table between Y and Z . We consider two estimators. One estimator is based on incomplete two-way stratification (analogous to the first set of constraints of Section 2.2), and the other estimator is based on a mix between statistical matching and calibration (analogous to the third set of constraints of Section 2.2).

3.1 Incomplete Two-Way Stratification

First the population totals of Y and Z are estimated by means of the first and second (large) sample respectively. These population totals are estimated in two phases. In the first phase, both (large) samples are weighted using X as a set of control variables. This implies that both (large) samples are weighted such that they reproduce the known population totals of X , which are denoted by t_x . Based on these weights, a pooled estimate for the population totals of U is

$$\hat{t}_u = \lambda \sum_{k \in n_1} w_{1k} u_k + (1 - \lambda) \sum_{k \in n_2} w_{2k} u_k,$$

where w_{1k} and w_{2k} denote the (first phase) calibration weights of the first and second sample, and $\lambda \in [0, 1]$. In

the second phase, both samples are reweighted using simultaneously X and U as control variables. Let v_{1k} and v_{2k} denote these second phase calibration weights. The resulting estimators for the population totals of Y and Z can be considered as calibration estimators in two phases (see Renssen and Nieuwenbroek 1997, Section 6). These estimators are denoted by \hat{t}_y and \hat{t}_z respectively:

$$\hat{t}_y = \sum_{k \in n_1} v_{1k} y_k \text{ and } \hat{t}_z = \sum_{k \in n_2} v_{2k} z_k.$$

We note that both estimators are based on a similar set of control variables. If the common set of variables is large, one may consider using a smaller subset to weight both samples. In general, the subset to weight the first sample may differ from the subset to weight the second sample. However, we shall assume in the sequel that both (large) samples are weighted according to the same set of control variables.

The two-way table between Y and Z can be estimated by weighting the (small) third sample, using simultaneously Y and Z as control variables, *i.e.*,

$$\hat{T} = \sum_{k \in n_3} w_{3k} (y_k z_k^t),$$

where the calibration weights w_{3k} satisfy the constraints

$$\sum_{k \in n_3} w_{3k} y_k = \hat{t}_y \text{ and } \sum_{k \in n_3} w_{3k} z_k = \hat{t}_z.$$

This is incomplete two-way stratification, where the unknown population totals of Y and Z are replaced by their estimates. These sets of constraints ensure precisely estimated marginal counts of the YZ -table if the common variables C are highly correlated with Y and Z .

3.2 Synthetic Two-Way Stratification

In this section, we consider an alternative estimator for the YZ -table, which also uses the (large) samples as a source of auxiliary information. However, instead of using estimated marginal counts as auxiliary information, estimated synthetic cell counts are used. Let B denote the population regression coefficient between Y and C , which is estimated by the first (large) sample:

$$\hat{B} = \left(\sum_{k \in n_1} v_{1k} c_k c_k^t \right)^{-1} \left(\sum_{k \in n_1} v_{1k} c_k y_k^t \right).$$

Similarly, let A denote the population regression coefficient between Z and C , which is estimated by the second (large) sample:

$$\hat{A} = \left(\sum_{k \in n_2} v_{2k} c_k c_k^t \right)^{-1} \left(\sum_{k \in n_2} v_{2k} c_k z_k^t \right).$$

Note that these estimated regression coefficients are based on the second phase calibration weights instead of the inclusion weights. If there exists a constant a , such that $a'c_k = 1$ for all k , then we still have $l'\hat{B}'c_k = l'\hat{A}'c_k = 1$ for all k . Now, inspired by the decomposition given by (3), i.e.,

$$\sum_{k=1}^N y_k z_k' = B' \sum_{k=1}^N (c_k c_k') A + \sum_{k=1}^N (y_k - B'c_k)(z_k - A'c_k)',$$

we suggest estimating the two-way table in two steps. In the first step the first term on the right-hand side is estimated by substituting the population regression coefficients B and A by their estimates \hat{B} and \hat{A} . Furthermore, we suggest to estimate $\sum_c = \sum_{k=1}^N c_k c_k'$ by the pooled estimate:

$$\hat{\sum}_c = \gamma \sum_{k \in n_1} v_{1k} (c_k c_k') + (1 - \gamma) \sum_{k \in n_2} v_{2k} (c_k c_k'),$$

where v_{1k} and v_{2k} denote the (second phase) weights of the first and second sample and $\gamma \in [0, 1]$. Eventually, the first term is estimated by $\hat{B}' \hat{\sum}_c \hat{A}$. Until now, no use of the third (small) sample has been made. If desired, estimates for B , A , and \sum_c can be improved slightly by also using the small sample.

In the second step, the complete two-way table between Y and Z is estimated by weighting the third (small) sample according to the calibration estimator subject to the third set of constraints (see Section 2.2), where B , A , and \sum_c are replaced by their estimates \hat{B} , \hat{A} , and $\hat{\sum}_c$. The resulting estimator equals

$$\sum_{k=1}^{n_3} w_{3k} (y_k z_k') = \hat{B}' \hat{\sum}_c \hat{A} + \sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)'. \quad (8)$$

The first term on the right-hand side is an estimate for the synthetic two-way table. This estimate is approximately unbiased for the YZ -table, if the conditional independence assumption holds. We note that, this type of estimator is essentially obtained by applying the constrained statistical matching method (see e.g., Barr and Turner 1980, Rodgers 1984, or Rubin 1986). The second term is an adjustment term to obtain an approximately unbiased estimate for the YZ -table, without this assumption. If there exists a constant a such that $a'c_k = 1$ for all sampled elements, then we obtain by pre-multiplying both sides of (8) with l' , the following estimator for the population total of Z :

$$\sum_{k \in n_3} w_{3k} z_k' = \left(\gamma \sum_{k \in n_1} v_{1k} c_k' + (1 - \gamma) \sum_{k \in n_2} v_{2k} c_k' \right) \hat{A} = \left(\sum_{k \in n_2} v_{2k} c_k' \right) \hat{A} = a' \left(\sum_{k \in n_2} v_{2k} c_k c_k' \right) \hat{A} = \hat{t}_z'.$$

Similarly, we have by post-multiplying both sides with l , an estimator for the population total of Y :

$$\sum_{k \in n_3} w_{3k} y_k = \hat{B}' \left(\gamma \sum_{k \in n_1} v_{1k} c_k + (1 - \gamma) \sum_{k \in n_2} v_{2k} c_k \right) = \hat{B}' \left(\sum_{k \in n_1} v_{1k} c_k \right) \hat{A} = \hat{t}_y'.$$

It follows that the marginal cell counts of the estimated two-way table, are the two-phase calibration estimators for the population totals of Y and Z as defined Section 3.1.

3.3 A Simulation Study; Integration of Household Surveys

In this subsection, we wish to compare the weighting techniques incomplete two-way stratification as discussed in subsection 3.1, and synthetic two-way stratification as discussed in subsection 3.2, by means of a simulation study. To that purpose, we use a data set, which stems from a pilot study of the Dutch Household Survey on Living Conditions, (see van Tuinen 1995). The data set consists of 1,085 records of which the following variables are observed: age (six categories: 15-24, 25-34, 35-44, 45-54, 55-64, 65+), sex (two categories: male or female), ownership of house (two categories: yes or no), occupation (five categories: work, housekeeping, education, voluntary, other), and health (two categories: yes or no). On behalf of the simulation study, this data set is considered as a finite population. The population totals of age and sex are assumed to be known.

In order to simulate the weighting techniques, we have carried out a Monte Carlo algorithm. Namely, we have drawn 500 samples, independently of each other, according to a two-phase sampling design. In the first phase, a simple random sample of size 20,500 is drawn with replacement. In this sample, age, sex, and ownership of house, are observed. In the second phase, the (first phase) sample is randomly divided into two large sub-samples of sizes 10,000 and one small sub-sample of size 500; in the one large sub-sample, occupation is observed (denoted by Y), in the other large sub-sample, health (denoted by Z), and in the small sub-sample, both occupation and health are observed. At each run, we have estimated the two-way table between Y and Z , according to four weighting methods which are discussed next.

The first phase sample is weighted with a crossing between sex and age as control variables. This is just post-stratification with twelve post-strata. Based on these weights, population totals can be estimated for all observed variables in the first phase sample, and for crossings between them. In particular, we may reproduce the population totals for the crossing between age and sex, and obtain estimated population totals for the crossings between age, sex, and ownership of house. Now, we distinguish two sets of common variables to weight the large sub-samples, as well as to obtain an estimate for the synthetic two-way table between Y and Z . The first set is a crossing between age and sex (12 categories) and the second set is a crossing between age, sex, and ownership (24 categories). For each simulation, this gives two different estimates for the marginal counts, *i.e.* two different estimates for the population totals of Y and Z – note that both estimates are based on post-stratification – and two different estimates for the synthetic two-way table. In order to weight the small sub-sample, we distinguish between the weighting method based on incomplete two-way stratification, and the weighting method based on synthetic two-way stratification. Since two different sets of common variables are used to weight the large sub-samples, as well as for statistical matching, we obtain four sets of calibration weights for each simulation run with respect to the small sub-sample, which in turn gives for each simulation run, four different estimated two-way tables between Y and Z . For the ease of computation, we have used the quadratic distance measure in the calibration estimation, implying that each estimated cell corresponds to a general regression estimate. Finally, we have taken the averages and variances of these two-way tables over the 500 simulations. The results are shown in tables 5 to 8.

The averages over the 500 simulations are almost identical for the four types of estimators, as can be seen from these tables. Note that the given cell counts are rounded off. We have also calculated the real YZ -table from the finite population. The real counts equal exactly the averages, which are given in Table 5 (or 6). For this particular simulation study, we conclude that all estimators have a very small bias.

The variances over these 500 simulations are given within the brackets. The variances of the estimated marginal counts of Tables 5 and 7 coincide, because these estimates are based on the same estimator. For the same reason it holds that the variances of the estimated marginal counts in tables 6 and 8 coincide. Note that the variances of the estimated marginal counts in tables 6 and 8 are slightly smaller than the variances of the estimated marginal counts in Tables 5 and 7, due to the larger set of common variables. However, for most estimated marginal counts this variance reduction can be considered negligible.

Tables 5 and 6 give identical variances with respect to all estimated cell counts. The variances for most estimated cell counts in Table 7, are plainly smaller than those in tables 5

and 6. In Table 8, this variance reduction is even greater. For this particular example, we conclude that the use of the larger set of common variables, in combination with the first weighting method, slightly reduces the variances of the estimated marginal counts, but leaves the variances of the estimated cell counts unaffected. Naturally, using the larger set of common variables in combination with the second weighting method, also slightly reduces the variances of the marginal cell counts. Finally, given a set of common variables, the weighting method based on synthetic matching, results in smaller variances for the estimated cell counts, than the weighting method based on incomplete two-way stratification.

Table 5
Incomplete Two-way Stratification Combined with the First Set of Common Variables

	1	2	3	4	5	total
yes	447 ₍₉₆₎	232 ₍₉₇₎	89 ₍₂₈₎	25 ₍₂₁₎	59 ₍₄₉₎	852 ₍₁₇₎
no	61 ₍₇₉₎	104 ₍₉₀₎	11 ₍₂₁₎	11 ₍₁₉₎	46 ₍₄₆₎	233 ₍₁₇₎
total	508 ₍₂₃₎	336 ₍₁₉₎	100 ₍₈₎	36 ₍₃₎	105 ₍₁₀₎	1085

Table 6
Incomplete Two-way Stratification Combined with the Second Set of Common Variables

	1	2	3	4	5	total
yes	447 ₍₉₆₎	232 ₍₉₇₎	89 ₍₂₈₎	25 ₍₂₁₎	59 ₍₄₉₎	852 ₍₁₇₎
no	61 ₍₇₉₎	104 ₍₉₀₎	11 ₍₂₁₎	11 ₍₁₉₎	46 ₍₄₆₎	233 ₍₁₇₎
total	508 ₍₂₃₎	336 ₍₁₉₎	100 ₍₈₎	36 ₍₃₎	105 ₍₉₎	1085

Table 7
Synthetic Two-way Stratification Combined with the First Set of Common Variables

	1	2	3	4	5	total
yes	447 ₍₇₅₎	231 ₍₇₄₎	89 ₍₁₇₎	25 ₍₂₀₎	59 ₍₄₂₎	851 ₍₁₇₎
no	61 ₍₅₈₎	105 ₍₆₅₎	11 ₍₁₂₎	11 ₍₁₉₎	46 ₍₃₈₎	234 ₍₁₇₎
total	508 ₍₂₃₎	336 ₍₁₉₎	100 ₍₈₎	36 ₍₃₎	105 ₍₁₀₎	1085

Table 8
Synthetic Two-way Stratification Combined with the Second Set of Common Variables

	1	2	3	4	5	total
yes	447 ₍₇₀₎	231 ₍₇₀₎	89 ₍₁₆₎	25 ₍₁₈₎	59 ₍₄₀₎	851 ₍₁₇₎
no	61 ₍₅₂₎	105 ₍₆₀₎	11 ₍₁₁₎	11 ₍₁₆₎	46 ₍₃₇₎	234 ₍₁₇₎
total	508 ₍₂₃₎	336 ₍₁₉₎	100 ₍₈₎	36 ₍₃₎	105 ₍₉₎	1085

3.4 Imputing Values of the one Large Sample into the Other Large Sample

By means of the two large samples and the small sample, one may construct a synthetic sample in which the real Y -values and predicted Z -values, and/or the predicted Y -values and the real Z -values are simultaneously recorded.

We define predictions for the Y - and Z -values analogously to (7), namely

$$\hat{y}_k = \hat{B}' c_k + \tilde{\beta}_2' (z_k - \hat{A}' c_k), k = 1, \dots, n_2, \quad (9)$$

and

$$\hat{z}_k = \hat{A}' c_k + \tilde{\alpha}_2' (y_k - \hat{B}' c_k), k = 1, \dots, n_1, \quad (10)$$

with

$$\tilde{\beta}_2 = \left[\sum_{k=1}^{n_2} v_{2k} (z_k - \hat{A}' c_k) (z_k - \hat{A}' c_k)' \right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}' c_k) (z_k - \hat{A}' c_k)' \right],$$

and

$$\tilde{\alpha}_2 = \left[\sum_{k=1}^{n_1} v_{1k} (y_k - \hat{B}' c_k) (y_k - \hat{B}' c_k)' \right]^{-1} \times \left[\sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}' c_k) (z_k - \hat{A}' c_k)' \right].$$

For each (c_k, y_k) the Z -values can be imputed in the first large sample by means of (10), $k = 1, \dots, n_1$, and similarly for each (c_k, z_k) the Y -values can be imputed in the second large sample by means of (9), $k = 1, \dots, n_2$. Based on these imputed values, we may define the following estimates for the two-way table between Y and Z :

$$\sum_{k=1}^{n_1} v_{1k} y_k \hat{z}_k' = \hat{B}' \sum_{k=1}^{n_1} v_{1k} c_k c_k' \hat{A} + \sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}' c_k) (z_k - \hat{A}' c_k)' \quad (11)$$

and

$$\sum_{k=1}^{n_2} v_{2k} \hat{y}_k z_k' = \hat{B}' \sum_{k=1}^{n_2} v_{2k} c_k c_k' \hat{A} + \sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}' c_k) (z_k - \hat{A}' c_k)' \quad (12)$$

One estimate is based on the first synthetic sample, the other on the second synthetic sample. By pooling the synthetic samples, one obtains a pooled synthetic sample of size $n_1 + n_2$, from which a pooled estimated for the two-way table can be constructed. This pooled estimate shows a close resemblance to (8). Note that if C and Z are perfectly correlated, then the left-hand side of (11) reduces to $\sum_{k=1}^{n_1} v_{1k} y_k z_k'$, i.e., our estimated two-way table corres-

ponds to a weighted estimated two-way table based on the first sample, as if the real values of Z were imputed in this sample. Similarly, if C and Y are perfectly correlated, then (12) reduces to $\sum_{k=1}^{n_2} v_{2k} y_k z_k'$.

An important special case to consider, is when c is categorical. Then the following equalities hold true:

$$\sum_{k \in n_1} v_{1k} (c_k c_k') = \sum_{k \in n_2} v_{2k} (c_k c_k') = \text{diag} \begin{pmatrix} t_x \\ t_u \end{pmatrix},$$

so (11) and (12) coincide. Furthermore, we have for categorical c :

$$\sum_{k \in n_1} v_{1k} c_k \hat{z}_k' = \sum_{k \in n_2} v_{2k} c_k z_k'$$

and

$$\sum_{k \in n_2} v_{2k} \hat{y}_k c_k' = \sum_{k \in n_1} v_{1k} y_k c_k'.$$

Obviously, if c is categorical, then it suffices to create a synthetic sample, which is based on either the first synthetic sample or the second synthetic sample. In either case, the weighting type estimates for the CZ -table, the CY -table, and the YZ -table, can be reconstructed. Finally, we note that the imputed values in all synthetic samples may be unrealistic. As described in Section 2.4, the calculated predictions may be replaced by live values according to some algorithm.

4. SUMMARY

In this article we presented a weighting procedure to combine information from distinct sample surveys. The linking pin between these surveys, is a set of common variables, (see Figure 1). It is argued that these samples should be weighted according to a sequential structure. First, both large samples were weighted using X as control variables. Based on these weighted samples, we could obtain a pooled estimate for the population total of U . Then both large samples were reweighted using simultaneously X and U as control variables. This gave an estimate for the population total of Y and Z .

Using statistical matching techniques with X and U as common variables, we may also obtain an estimate for a synthetic two-way table between Y and Z . Eventually, the small sample was weighted according to two different sets of control variables. The first set of control variables corresponded to the estimated population totals of Y and Z , and the second set of control variables to the estimated synthetic two-way table. Using the first set of control variables, is strongly related to incomplete two-way stratification. The theoretical framework needed to develop the second weighting method, was discussed all through this article. By means of both weighting methods, the

YZ-table can be estimated (it is tacitly assumed that Y and Z are categorical). The marginal counts of the YZ-table corresponding to the first weighting method, equal by definition of the calibration equations, the estimated population totals of Y (which is based on the first large sample) and Z (which is based on the second large sample). It was shown, that this consistency property also holds for the second weighting method. A numerical study was conducted to evaluate the performance of the weighting methods with respect to the cell counts. It was found that both weighting methods yielded nearly (design) unbiased estimated two-way tables. The simulated (design) variances of the second weighting method, appeared to be smaller than the corresponding (design) variances of the first weighting method, with respect to all estimated cell counts. In principle, the Y - and Z -variables were assumed to be categorical, however, it was argued that the ideas presented were also applicable for continuous Y and Z or for continuous Y and categorical Z .

ACKNOWLEDGEMENTS

The author wishes to thank Peter Kooiman, Nico Nieuwenbroek, and Ger Slootbeek for their careful reading and useful remarks. The author also thanks two anonymous referees and an associated editor for their valuable suggestions to improve the article. The views expressed in this article are those of the author and do not necessarily reflect the policy of Statistics Netherlands.

REFERENCES

- ALEXANDER, C.H. (1987). A class of methods for using person controls in household weighting. *Survey Methodology*, 13, 183-198.
- BAKKER, B.F.M., and WINKELS, J.W. (1998). Why integration of household surveys? – Why POLS?. *Netherlands Official Statistics*, 13, 5-7.
- BARR, R.S., and TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.
- BETHLEHEM, J.G., and KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- COPELAND, K.R., PEITZMEIER, F.K., and HOY, C.E. (1987). An alternative method of controlling Current Population Survey estimates to population counts. *Survey Methodology*, 13, 173-181.
- DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.C., SÄRNDAL, C.-E., and SAUTORY, O., (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- FELLEGI, I.P., and SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- HOFMANS, M.G. (1998). Innovative weighting in POLS. Making use of core questions. *Netherlands Official Statistics*, 13, 12-15.
- LEMAÎTRE, G., and DUFOUR, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13, 199-208.
- MADDALA, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- OH, H.L., and SCHEUREN, F. (1987). Modified raking ratio estimation. *Survey Methodology*, 13, 209-219.
- PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. In *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science.
- RAGHUNATHAN, T.E., and GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.
- RENSSEN, R.H., and NIEUWENBROEK, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.
- RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2, 91-102.
- RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 87-94.
- SÄRNDAL, C.-E., SWENSSON, B., and WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons.
- SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons.
- SINGH, A.C., MANTEL, H.J., KINACK, M.D., and ROWE, R. (1993). Statistical matching: use of auxiliary information as an alternative to the conditional independence assumption. *Survey Methodology*, 19, 59-79.
- TUINEN VAN, H.K. (1995). Social indicators, social surveys and integration of social statistics. *Statistical Journal of the United Nations ECE*, 12, 379-394.
- WINKELS, J.W., and EVERAERS, P.C.J. (1998). Design of an integrated survey in the Netherlands. The case POLS. *Netherlands Official Statistics*, 13, 6-11.
- ZIESCHANG, K.D. (1990). A generalized least squares weighting system for the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

Sampling on Two Occasions: Estimation of Population Total

RAGHUNATH ARNAB¹

ABSTRACT

Two sampling strategies have been proposed for estimating the finite population total for the most recent occasion, based on the samples selected over two occasions involving varying probability sampling schemes. Attempts have been made to utilize the data collected on a study variable, in the first occasion, as a measure of size and a stratification variable for selection of the matched-sample on the second occasion. Relative efficiencies of the proposed strategies have been compared with suitable alternatives.

KEY WORDS: Composite estimator; Matched-sample; Sampling schemes; Sampling strategies; Varying probability sampling schemes.

1. INTRODUCTION

We very often survey the same population at regular time intervals to estimate the same population characteristics which change over time. For example, many countries collect data to estimate total number of unemployed persons, HIV infected people, immigrants *etc.*, on an annual or quarterly basis. In this article, we consider a finite population $U = (U_1, \dots, U_i, \dots, U_N)$ of N identifiable units, which is supposed to be sampled over two occasions, to estimate the population total of a variable under study for the current (second) occasion. In successive sampling, one utilizes data collected on the previous (first) occasion effectively, to get an efficient strategy in consideration of cost, and providing an efficient estimator of the population total for the current occasion. Extensive literature is now available for this purpose. Singh (1967), and Avadhani and Sukhatme (1970) utilized information, collected on the first occasion as a measure of size, for the selection of the matched sample on the second occasion; while Arnab (1991) utilized such information as a stratification variable, as well as the measure of size, for selection of the sample on the second occasion. Recently, Prasad and Graham (1994) modified Raj's (1965) and Chotai's (1974) sampling strategies, by using information of the first occasion as a measure of size, for the selection of the matched sample in the second occasion. They found empirically, that one of their proposed strategies fares better than that given by Chotai (1974). In this article, two alternative strategies are proposed. One of them utilizes information in the first occasion as a measure of size, and the other utilizes information as a measure of size and also as a stratification variable for selection of the matched sample in the second occasion. In this paper, it is shown that one of the proposed strategies is better than that given by Prasad and Graham (1994) and for the other, we do not have any definite theoretical conclusion. However, empirical evidence shows that the latter is more efficient

than that described by Prasad and Graham (1994), as well as the former proposed strategy. This is possible because it utilizes first occasion values in all possible stages *viz.*, stratification, estimation and selection of the matched sample in the second occasion.

The general methods of selection of samples and estimation over two occasions are described below.

1.1 Sampling Schemes

On the first occasion, a sample s_1 , of size n , is selected by some suitable sampling design, say P_1 , and the data y_{1i} , $i \in s_1$, is obtained where y_{1i} (y_{2i}) is the value of the variate y under study, for the i -th unit on the first (second) occasion. On the second occasion, a matched sample (sub-sample) s_m of size $m (= n\lambda)$, assumed to be an integer, $0 \leq \lambda \leq 1$ is selected from s_1 by some suitable sampling scheme P_m , and it is supplemented by an un-matched sample s_u of size $u (= n\mu = n - m, \mu = 1 - \lambda)$ either from the entire population U or from U/s_1 , the set of units not selected in the first occasion, by some suitable sampling design P_u , and information y_{2i} ($i \in s_m, i \in s_u$) on the second occasion is obtained. It is obvious that the cost of survey for the matched sampled units is expected to be much lower than that of the un-matched units, but for the sake of simplicity, we assume that the cost of the survey remains the same for all the units in the second occasion.

1.2 Method of Estimation

From the data y_{1i} , $i \in s_1$, and y_{2i} , $i \in s_m$ collected through the initial sample s_1 , and the matched sample s_m , an unbiased estimator \hat{Y}_{2m} for Y_2 , the population total for the second occasion, is formed by treating the y_{1i} 's, $i \in s_1$, as auxiliary information. Thus \hat{Y}_{2m} is normally a difference, ratio or regression estimator. From the un-matched sample s_u , an unbiased estimator \hat{Y}_{2u} is also constructed for Y_2 . Finally, a composite estimator, a combination of \hat{Y}_{2m} and

¹ Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa.

\hat{Y}_{2u} , is obtained by using a suitable weight of ϕ ($0 \leq \phi \leq 1$), as

$$\hat{Y}_2 = \phi \hat{Y}_{2m} + (1 - \phi) \hat{Y}_{2u}. \quad (1)$$

The optimum value of $\phi = \phi(\lambda)$ is obtained by minimizing $V(\hat{Y}_2)$, the variance of \hat{Y}_2 with respect to ϕ , for a given value of m (i.e., λ). The expressions for $\phi(\lambda)$ and $V(\hat{Y}_2 I \lambda)$, the variance of \hat{Y}_2 with $\phi = \phi(\lambda)$ are obtained as follows, when \hat{Y}_{2m} and \hat{Y}_{2u} are independent:

$$\phi(\lambda) = (1/V_m) [1/V_m + 1/V_u]^{-1},$$

$$V(\hat{Y}_2 I \lambda) = [1/V_m + 1/V_u]^{-1},$$

where V_m and V_u are variances of \hat{Y}_{2m} and \hat{Y}_{2u} respectively. The optimum proportion of matched sample $\lambda = \lambda_0$, is obtained by minimizing $V(\hat{Y}_2 I \lambda)$ with respect to λ . Finally, putting $\lambda = \lambda_0$ in the expression for $V(\hat{Y}_2 I \lambda)$, the minimum variance of \hat{Y}_2 is obtained, and it will be denoted by $V_{\min}(\hat{Y}_2) = V(\hat{Y}_2 I \lambda_0)$. Our object is to find a suitable strategy, which is a combination of $P = (P_1, P_m, P_u)$ and \hat{Y}_2 , to control the magnitude of $V_{\min}(\hat{Y}_2)$ to a minimum.

1.3 A Few Sampling Strategies

1.3.1 Avadhani and Sukhatme (1970)

On the first occasion, the initial sample s_1 of size n was selected by simple random sampling without replacement (SRSWOR) method, assuming that no auxiliary information is available prior to this survey. On the second occasion, the matched sample s_m of size m was selected from s_1 by the Rao, Hartley and Cochran (RHC, in brief, 1962) sampling scheme using y_{1i} as a measure of size for the i -th unit $i \in s_1$, assuming y_{1i} 's are positive. Under the RHC sampling scheme, the selected n units of s_1 , are divided at random into m groups, each of size n/m , which is assumed to be an integer. From each of the selected groups, one unit is selected independently with probability proportional to the measure of size. Thus if the i -th unit, U_i , belongs to the j -th group G_j ($j = 1, \dots, m$) then U_i will be selected with the probability $q_i^*(i \in s_1) = y_{1i} / \sum_{i \in s_1} y_{1i}$. The un-matched sample s_u was selected from U/s_1 by SRSWOR.

1.3.2 Chotai (1

On the first occasion, the initial sample s_1 of size n was selected by the RHC scheme of sampling (assuming N/n is an integer), as described above with probability proportional to z_i , the size measure for the i -th unit which is, assumed to be positive and known for every $i \in U$. Let $\Delta_j = \sum_{k \in G_j} p_k$, the sum of p_k ($= z_k/Z$, $Z = \sum_{i \in U} z_i$) values that belong to the random group G_j ($j = 1, \dots, n$), which is formed in selecting the sample s_1 by the RHC method. The matched sample s_m was selected from s_1 by the RHC

scheme, with normed size measure Δ_i , for the i -th unit $i \in s_1$ ($\sum_{i \in s_1} \Delta_i = 1$) assuming n/m is an integer. The un-matched sample, s_u was selected by the RHC sampling scheme with normed size measure p_i for the i -th unit assuming N/u is an integer. Let P_i^+ (P_i') = total of the Δ_i (p_i) values associated with those units that belong to the random group from which the i -th unit was selected in s_m (s_u) by the RHC sampling scheme with $\sum_{i \in s_m} P_i^+ = 1$ ($\sum_{i \in s_u} P_i' = 1$).

The composite estimator for Y_2 is given by

$$\hat{Y}_2 = \phi \hat{Y}_{2m} + (1 - \phi) \hat{Y}_{2u}$$

where

$$\begin{aligned} \hat{Y}_{2m} &= \sum_{i \in s_m} (y_{2i}/p_i) P_i^+ - \\ &\gamma \left[\sum_{i \in s_m} (y_{1i}/p_i) P_i^+ - \sum_{i \in s_1} (y_{1i}/p_i) \Delta_i \right]; \\ \hat{Y}_{2u} &= \sum_{i \in s_u} (y_{2i}/p_i) P_i' \end{aligned} \quad (2)$$

where γ is a suitably chosen constant to minimize variance of \hat{Y}_{2m} . Chotai (1974) derived the expression for the minimum variance of \hat{Y}_2 as

$$V_{\min}(\hat{Y}_2) = k [1 - f + \sqrt{(1 - \delta^*)}] \sigma_2^2 / 2 = V_c \text{ (say)} \quad (3)$$

where

$$k = N / \{n(N - 1)\}, f = n/N,$$

$$\sigma_t^2 = \sum_{i \in U} p_i (y_{ti}/p_i - Y_t)^2, t = 1, 2$$

$$Y_t = \sum_{i \in U} y_{ti}, t = 1, 2$$

$$\delta^* = \sum_{i \in U} p_i (y_{2i}/p_i - Y_2) (y_{1i}/p_i - Y_1) / (\sigma_1 \sigma_2). \quad (4)$$

1.3.3 Arnab (1991)

Arnab (1991) presented several strategies where the initial sample s_1 was selected by probability proportional to size with replacement (PPSWR) using normed size measure $p_i = z_i/Z$ for the i -th unit. Utilizing the ascertain values y_{1i} 's ($i \in s_1$) on the basis of certain criteria, the n sample units are assigned to a suitable number of L strata. Let s_{1h} be the sample of size n_h , belonging to the h -th stratum ($s_1 = \cup_h s_{1h}$ and $\sum_h n_h = n$). Here, it is assumed that n is large enough to ensure that n_h is positive for every h in practice. On the second occasion, sub-samples s_{mh} 's

of size m_h 's ($= v_h n_h$, v_h is a predetermined fraction and m_h is assumed to be an integer) are selected from s_{1h} 's independently, by suitable sampling schemes involving y_{1i} 's, $i \in s_1$ in the selection of matched samples s_{mh} 's. The unmatched sample s_u is selected by PPSWR method from the entire population U using z_i' as measure of size.

1.3.4 Prasad and Graham (1994)

Here the initial sample s_1 is selected by the RHC scheme of sampling similar to Chotai (1974) with normed size measure $p_i = z_i / Z$ for the i -th unit. The matched sample s_m is selected from s_1 by the RHC scheme with $p_i^* = (y_{1i} \Delta_i / p_i) / \sum_{i \in s_1} (y_{1i} \Delta_i / p_i)$ for the i -th unit, $i \in s_1$; where Δ_i is the sum of the p_j values for the group containing the i -th unit, formed in selecting s_1 by the RHC sampling scheme of sampling. The un-matched sample, s_u was selected from the entire population U by the RHC scheme similar to that presented by Chotai (1974). Here also N/n , n/m and N/u are assumed to be integers. Prasad and Graham (1994) proposed the following composite estimator for Y_2 :

$$\hat{Y}_2 = \phi \hat{Y}_{2m} + (1 - \phi) \hat{Y}_{2u}$$

where $\hat{Y}_{2m} = \sum_{i \in s_m} (y_{2i}^* / p_i^*) \tilde{P}_i$; $\hat{Y}_{2u} = \sum_{i \in s_u} (y_{2i} / p_i) P_i'$; $y_{2i}^* = y_{2i} \Delta_i / p_i$; $\tilde{P}_i (P_i')$ = total of the $p_i (p_i)$ values associated with those units that belong to the random group from which the i -th unit was selected in $s_m (s_u)$. The expression for minimum variance of \hat{Y}_2 , is obtained as:

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta}) \sigma_2^2 / 2 = V_{PG} \text{ (say)} \quad (5)$$

where

$$\zeta = \sigma_3^2 / \sigma_2^2, \sigma_3^2 = \sum_{i \in U} q_i (y_{2i} / q_i - Y_2)^2, q_i = y_{1i} / Y_1; \quad (6)$$

k, f, σ_2^2 and Y_1 are defined in (4).

In Prasad and Graham's (1994) expression for $V_{\min}(\hat{Y}_2)$, the divisor 2 was omitted and is obviously a typographical error.

Remark 1.1

From the strategies described in section 1.3, we note that the Avadhani and Sukhatme (1970) scheme does not require information on size measures in the whole frame, and hence is less demanding than the others. Chotai (1974) used the original size measures p_i in selection, but the first survey values y_{1i} 's, $i \in s_1$ were used additionally in estimation only. The use of additional information, p_i 's, for the selection of the initial sample s_1 will make Chotai's (1974) strategy more efficient than that of Avadhani and Sukhatme (1970). But to use the optimal estimator \hat{Y}_2 for the Avadhani and Sukhatme (1970) strategy, one needs to estimate ϕ , the only unknown parameter. However, in Chotai's (1974) strategy, both the parameters ϕ and γ have

to be estimated in order to use the optimum \hat{Y}_2 . Prasad and Graham (1994) used both these variables in the selection of the matched sample (hence automatically in the estimation) and showed empirically that their strategy fares better than that of Chotai (1974). In addition, to gain in efficiency, Prasad and Graham's (1994) strategy can be used in practice, because \hat{Y}_2 involves only one unknown parameter, ϕ . It should be noted that Arnab (1991) first introduced the principle of stratification using y_{1i} 's, $i \in s_1$ as a stratification variable. This should always be done in practice whenever the necessary information is available, particularly in the selection of large units with marked size differences of the type considered in the numerical examples in section 3. Arnab's (1991) strategy is expected to be more efficient than the preceding strategies, since it utilizes first occasion values for stratification in addition to estimation. However, the optimal estimator \hat{Y}_2 contains the several unknown parameters (for details see Arnab 1991) which may hinder the application of the strategy especially when the sample size is not large enough.

2. PROPOSED STRATEGIES

Here two sampling strategies have been proposed which are modifications of strategies proposed by Prasad and Graham (1994) and Arnab (1991), respectively.

2.1 Strategy 1

The sampling scheme for this strategy is the same as was considered by Prasad and Graham (1994), and described in section 1.3.4. Here, only the estimator based on the matched sample s_m , has been modified by introducing the original size measure into the estimation. The proposed modified estimator \hat{Y}_{2m}^* and the composite estimators for Y_2 are as follows:

$$\hat{Y}_{2m}^* = \sum_{i \in s_m} (y_{2i}^* / p_i^*) \tilde{P}_i - \beta \left[\sum_{i \in s_m} (z_i^* / p_i^*) \tilde{P}_i - Z \right] = \sum_{i \in s_m} (r_i^* / p_i^*) \tilde{P}_i + \beta Z$$

where $z_i^* = z_i \Delta_i / p_i$, $y_{2i}^* = y_{2i} \Delta_i / p_i$, $r_i^* = r_i \Delta_i / p_i$, $r_i = y_{2i} - \beta z_i$ and β is a suitably chosen constant to minimize variance of \hat{Y}_{2m}^* ; p_i^* , \tilde{P}_i and Δ_i are as described in the section 1.3.4;

$$\hat{Y}_2 = \phi \hat{Y}_{2m}^* + (1 - \phi) \hat{Y}_{2u}$$

where \hat{Y}_{2u} is given in (2).

Denoting $E_1(V_1)$ as unconditional expectation (variance) over selection of the sample s_1 , and $E_2(V_2)$ the conditional expectation (variance) over s_m when s is fixed, one gets the variance of \hat{Y}_{2m}^* for a given value of β , as

$$V(\hat{Y}_{2m}^* | \beta) = E_1 V_2(\hat{Y}_{2m}^* | \beta) + V_1 E_2(\hat{Y}_{2m}^* | \beta).$$

Following Prasad and Graham (1994), we obtain

$$E_1 V_2(\hat{Y}_{2m}^* I\beta) = k_1 \sigma_3^{*2}(\beta)$$

and

$$V_1 E_2(\hat{Y}_{2m}^*) = k(1-f) \sigma_2^2$$

where

$$k_1 = N(n-m)/\{nm(N-1)\};$$

$$\begin{aligned} \sigma_3^{*2}(\beta) &= \sum_{i \in U} q_i (r_i/q_i - R)^2 \\ &= \sigma_3^2 + \beta^2 \sigma_0^2 - 2\beta \sigma_0 \sigma_3 \delta; \\ R &= \sum_{i \in U} R_i = Y_2 - \beta Z, \delta = \sigma_{03}/(\sigma_0 \sigma_3), \\ \sigma_0^2 &= \sum_{i \in U} q_i (z_i/q_i - Z)^2, \\ \sigma_{03} &= \sum_{i \in U} q_i (y_{2i}/q_i - Y_2)(z_i/q_i - Z) \end{aligned} \quad (7)$$

σ_2^2, k and σ_3^2, q_i are as in (4) and (6), respectively. The optimum value of β that minimizes $V(\hat{Y}_{2m}^* I\beta)$ comes out as, $\text{opt } \beta = \beta_0 = \delta \sigma_3 / \sigma_0$.

Putting the optimum value of $\beta = \beta_0$ in the expression of $V(\hat{Y}_{2m}^* I\beta)$, we get the optimum value of

$$V(\hat{Y}_{2m}^* I\beta) = V(\hat{Y}_{2m}^* I\beta_0) = k[(1-f) + (1-\lambda)\zeta^*/\lambda] \sigma_2^2$$

where $\zeta^* = (1 - \delta^2)\zeta$; k, f and ζ are defined in (4) and (6) respectively.

The optimum variance of \hat{Y}_2 for a given value of λ is obtained by minimizing the variance of \hat{Y}_2 with respect to ϕ when $\beta = \beta_0$, and is given by

$$\begin{aligned} V_{\text{opt}}(\hat{Y}_2 I\lambda) &= [1/V(\hat{Y}_{2m}^* I\beta_0) + 1/(\hat{Y}_{2u})]^{-1} \\ &= [1/\{k(1-f) + (1-\lambda)\zeta^*/\lambda\} + \mu/\{k(1-f\mu)\}]^{-1} \sigma_2^2. \end{aligned}$$

Finally, minimizing $V_{\text{opt}}(\hat{Y}_2 I\lambda)$ with respect to λ , the optimum proportion of the matched sample and minimum variance of \hat{Y}_2 are obtained respectively as

$$\text{opt } \lambda = \lambda_0 = \sqrt{\zeta^*}/(1 + \sqrt{\zeta^*})$$

and

$$V_{\min}(\hat{Y}_2) = k(1-f + \sqrt{\zeta^*}) \sigma_2^2/2 = M_1 \text{ (say)} \quad (8)$$

Remark 2.1

The estimator \hat{Y}_{2m}^* , described in (1) is usable in practice when the optimum value of $\beta = \beta_0$ is known, or a good guess value of β_0 is available from some previous surveys. If instead of the regression estimator \hat{Y}_{2m}^* described above, one uses the difference estimator $\hat{Y}_{2m}^{**} = \sum_{i \in s_m} (y_{2i}/p_i^*) \tilde{P}_i - [\sum_{i \in s_m} (z_i/p_i^*) \tilde{P}_i - Z]$ based on the matched sample, the expression for the minimum variance of \hat{Y}_2 would be as follows:

$$V_{\min}(\hat{Y}_2) = k(1-f + \sqrt{\zeta}) \sigma_2^2/2 = \tilde{M}_1 \text{ (say)}$$

with

$$\tilde{\zeta} = (1 + \tau^2 - 2\tau\delta)\zeta, \tau = \sigma_0/\sigma_3.$$

2.1.1 Variance Estimation

To get approximate unbiased estimators for $V_{\text{opt}}(\hat{Y}_2)$, we first present the following theorems without proof:

Theorem 1

$$\begin{aligned} \hat{V}(\hat{Y}_{2m}^*) &= \{k/(1-k)\} \left\{ \left[\sum_{i \in s_m} (y_{2i}^2 \Delta_i / p_i^2) \tilde{P}_i / p_i^* - \hat{Y}_{2m}^{*2} \right] \right. \\ &\quad \left. + \{k_2/k\} \sum_{i \in s_m} \tilde{P}_i \left(r_i^* / p_i^* - \sum_{i \in s_m} \tilde{P}_i r_i^* / p_i^* \right)^2 \right\} \end{aligned}$$

is an unbiased estimator of $V(\hat{Y}_{2m}^*)$, when β_0 is known, $k = (N-n)/\{n(N-1)\}$ and $k_2 = (n-m)/\{m(n-1)\}$.

Theorem 2

$E_1 V_2 [\sum_{i \in s_m} \tilde{r}_i^* / p_i^*] = N(n-m)/\{nm(N-1)\} [\sigma_3^2 + \sigma_0^2 - 2\sigma_{03}]$ can be estimated unbiasedly by

$$\{(n-m)/n(m-1)\} \sum_{i \in s_m} (\tilde{r}_i^* / p_i^* - \sum_{i \in s_m} \tilde{r}_i^* / p_i^*)^2 \tilde{P}_i$$

where $\tilde{r}_i^* = \tilde{r}_i \Delta_i / p_i$, $\tilde{r}_i = y_{2i} - z_i$; σ_3^2, σ_0^2 and σ_{03} are given in (4) and (7) respectively.

From the Theorem 2 we note that

$$\hat{\sigma}_0^2 = d \sum_{i \in s_m} \left(z_i / p_i^* - \sum_{i \in s_m} z_i \tilde{P}_i / p_i^* \right)^2 \tilde{P}_i,$$

$$\hat{\sigma}_3^2 = d \sum_{i \in s_m} \left(y_{2i} / p_i^* - \sum_{i \in s_m} y_{2i} \tilde{P}_i / p_i^* \right)^2 \tilde{P}_i$$

and

$$\hat{\sigma}_{30}^2 = d \sum_{i \in s_m} \left(z_i / p_i^* - \sum_{i \in s_m} z_i \tilde{p}_i / p_i^* \right) \left(y_{2i} / p_i^* - \sum_{i \in s_m} y_{2i} \tilde{p}_i / p_i^* \right) \tilde{p}_i$$

are unbiased estimators of σ_0^2 , σ_3^2 and σ_{30}^2 , respectively where $d = m(N-1)/\{N(m-1)\}$.

Estimator for $V_{\text{opt}}(\hat{Y}_2 I \lambda)$

Thus for a given value of m (i.e., λ), we can suggest an approximate unbiased estimator of $V_{\text{opt}}(\hat{Y}_2 I \lambda)$ as,

$$V_{\text{opt}}(\hat{Y}_2 I \lambda) = (1/\hat{V}_m + 1/\hat{V}_u)^{-1},$$

where $\hat{V}_m = \hat{V}(\hat{Y}_{2m} I \beta_0)$ and \hat{V}_u is an unbiased estimator of $V(\hat{Y}_{2u}) = \{(N-u)/N(u-1)\} \sum_{i \in s_u} P_i' (y_{2i}/p_i - \hat{Y}_{2u})^2$.

Estimator for $V_{\min}(\hat{Y}_2)$

Putting suitable estimators for λ, ζ^* and σ_2^2 in the expression for $V_{\min}(\hat{Y}_2)$, we get an approximate unbiased estimator for $V_{\min}(\hat{Y}_2)$ as,

$$\hat{V}_{\min}(\hat{Y}_2) = k[1 - f + (1 - \hat{\lambda}) \hat{\zeta}^* / \hat{\lambda}] / \hat{\sigma}_2^2,$$

where

$$\hat{\zeta}^* = (1 - \delta^2) \hat{\zeta}, \hat{\lambda} = \sqrt{\hat{\zeta}^*} / (1 + \sqrt{\hat{\zeta}^*}),$$

$$\hat{\delta} = \hat{\sigma}_{03} / (\hat{\sigma}_0^2 \hat{\sigma}_3^2)^{1/2}, \hat{\zeta}^* = \hat{\sigma}_3^2 / \hat{\sigma}_2^2,$$

$$\hat{\sigma}_2^2 = \hat{\lambda} \hat{\sigma}_2^2(m) + (1 - \hat{\lambda}) \hat{\sigma}_2^2(u)$$

$\hat{\sigma}_2^2(m)$ = an approximate unbiased estimator of σ_2^2 based on the matched sample $s_m = \sum_{i \in s_m} (y_{2i} \Delta_i / p_i^2) \tilde{p}_i / p_i^* - \{\hat{Y}_{2m}^2 - \hat{V}_m\}$, $\hat{\sigma}_2^2(u)$ = an approximate unbiased estimator of σ_2^2 based on the un-matched sample $s_u = u(N-1)/\{N(u-1)\} \sum_{i \in s_u} P_i' (y_{2i} P_i' / p_i - \hat{Y}_{2u})^2$; k and f are as in (4).

Remark 2.2

Ideally one should estimate σ_2^2 through the optimum combination of $\hat{\sigma}_2^2(m)$ and $\hat{\sigma}_2^2(u)$ and in this case, the optimum combination will involve unknown parameters. To avoid this complexity, the simpler estimator ($\hat{\sigma}_2^2$) of σ_2^2 has been suggested above.

2.2. Strategy 2

The population is supposed to consist of L strata with N_h as the unknown size of the h -th stratum ($h = 1, \dots, L$; $\sum_h N_h = N$) stipulating that one can identify the stratum to which a unit belongs, as soon as its value is observed on the first occasion. On the first occasion, the initial sample s_1 of size n was selected by PPSWR method with normed size p_i

attached to the i -th unit. Let n_h units of s_1 , falling in the h -th stratum, be denoted as s_{1h} . Let $y_{1i}(h), y_{2i}(h)$ be respectively the value of the variate under study, of the i -th unit of the h -th stratum for the first and second occasions, and $z_i(h)$ be the corresponding size measure. On the second occasion, independent samples s_{mh} 's of sizes $m_h = m n_h / n$ (assumed an integer for every h), keeping $\sum_h m_h = m$ as fixed, are selected by the RHC sampling scheme with normed size $q_{hi}^* = [y_{1i}(h)/z_i(h)] / \sum_{i \in s_1} [y_{1i}(h)/z_i(h)]$ for the i -th unit of h -th stratum. The unmatched sample s_u was selected from the entire population by the RHC method with normed size measure p_i for the i -th unit as in strategy 1. The proposed estimators for Y_2 , based on the matched-sample s_m , and the un-matched sample s_u are respectively as follows:

$$\hat{Y}_{2m} = \sum_h w_h \hat{Y}_{2m}(h); \hat{Y}_{2u} = \sum_{s_u} (y_{2i}/p_i) P_i' \quad (9)$$

where

$$\hat{Y}_{2m}(h) = \sum_{s_{mh}} r_i(h) Q_{hi} / (n_{1h} p_{hi} q_{hi}^* + c_h \sum_{s_{1h}} z_j(h) /$$

$$(n_{1h} p_{hj}), w_h = n_{1h} / n, p_{hj} = z_j(h) / Z,$$

$$r_i(h) = y_{2i}(h) - c_h y_{1i}(h),$$

Q_{hi} = sum of q_{hj}^* for the group containing i -th unit of the h -th stratum, that was formed for selection of the matched sample s_{mh} by RHC method. c_h 's are constants chosen to minimize variance of $\hat{Y}_{2m}(h)$. Following Arnab (1991), the expression for variance of \hat{Y}_{2m} is obtained as:

$$V(\hat{Y}_{2m}) = k_2 \sum_h \sum_{j=1}^{N_h} q_{hj} (r_{hj} / q_{hj} - R_h)^2 / P(h) + \sigma_2^2 / n$$

where $k_2 = (n - m) / n$, $q_{hj} = y_{1j}(h) / y_1(h)$, $Y_1(h) = \sum_{j=1}^{N_h} y_{1j}(h)$, N_h = population size of the h -th stratum, $P(h) = Z_h / Z$, $Z = \sum_{j=1}^{N_h} z_j(h)$.

The optimum value of c_h that minimizes $V(\hat{Y}_{2m})$ and the corresponding value of $V(\hat{Y}_{2m})$ comes out respectively as

$$\text{opt } c_h = c_h(0) = \delta_{h3} = \sum_{j=1}^{N_h} q_{hj} \alpha_{hj} \beta_{hj} / (\sigma_{h0} \sigma_{h3})$$

and $[1 + (n - m)\theta/m] \sigma_2^2 / n$, where

$$\alpha_{hj} = y_{2j}(h) / q_{hj} - Y_2(h), \beta_{hj} = z_{hj} / q_{hj} - Z_h,$$

$$\sigma_{h3}^2 = \sum_{j=1}^{N_h} q_{hj} \alpha_{hj}^2, \sigma_{h0}^2 = \sum_{j=1}^{N_h} q_{hj} \beta_{hj}^2, Y_2(h) = \sum_{j=1}^{N_h} y_{2j}(h)$$

and $\theta = \sum_h (1 - \delta_h^2) \sigma_{h3}^2 / \{P_h \sigma_2^2\}$.

The proposed composite estimator for Y_2 , the optimum proportion of matched sample and the expression for the minimum variance of the composite estimator \hat{Y}_2 are given respectively by

$$\hat{Y}_2 = \phi \hat{Y}_{2m} + (1 - \phi) \hat{Y}_{2u}$$

$$\text{opt } \lambda = \lambda_0 = [\theta - (1 - f)\sqrt{\theta} \sqrt{f^*}] / [\theta + f\sqrt{\theta} \sqrt{f^*} - 1]$$

$$V_{\min}(\hat{Y}_2) = k(1/\mu_0 - f)\sigma_2^2 / [1 + (\lambda_0/\mu_0)\sqrt{f^*}/\sqrt{\theta}]$$

$$= M_2 \text{ (say)}$$

where \hat{Y}_{2m} and \hat{Y}_{2u} are given in (9), $f^* = N/(N - 1)$, $\mu_0 = 1 - \lambda_0$; k, f and σ_2^2 are given in (4).

3. EFFICIENCIES OF THE PROPOSED STRATEGIES

The proposed Strategy 1 is more efficient than the strategy proposed by Prasad and Graham (1994) in the sense of yielding smaller minimum variance, as $\delta^2 \leq 1$. Efficiency of the Strategy 1 increases as δ , the correlation between y_{2i}/q_i and z_i/q_i increases. The efficiency of the Strategy 1 and Prasad and Graham's (1994) strategy increases as ζ decreases. The value of $\zeta = \sigma_3^2/\sigma_2^2$ depends on the magnitudes of σ_3^2 and σ_2^2 . σ_3^2 will be smaller (greater) than σ_2^2 if the proportionality of y_{2i} on y_{1i} is higher (lower) than that of y_{2i} on z_i . Obviously, Strategy 1 can be used in practice when a good guess value of β is available from the past surveys. If the difference estimator is used in Strategy 1 instead of the regression estimator mentioned in Remark 2.1, then the proposed Strategy 1 fares better than that of Prasad and Graham (1994) whenever $\delta > \frac{1}{2}\sigma_0/\sigma_3$. Strategy 1 fares better or worse than Chotai's (1974) strategy according to $\zeta^* = (1 - \delta^2)\zeta < \text{or } > (1 - \delta^2)$. Here, δ^* may be regarded as a correlation coefficient between y_{2i}/p_i and y_{1i}/p_i . In particular, if z_i 's, are constant, then δ^* becomes the simple correlation coefficient between y_{1i} 's and y_{2i} 's. The expression for the minimum variance M_2 for Strategy 2 is complex and does not yield any simple comparison with the other strategies described here. However, we note that the efficiency of the Strategy 2 increases as the stratum correlation δ_{h3} increases. Following numerical examples based on the live data reveals that the proposed Strategy 2 fares better than Strategy 1 and also the alternatives proposed by Prasad and Graham (1994) and Chotai (1974).

For numerical comparisons, three data sets are considered. One of them (will be called Population 1) was considered by Prasad and Graham (1994) which relates to the area under wheat in 1937 (y_2) and 1936 (y_1) and cultivated area (z) for a set of 34 villages in India, compiled by Sukhatme and Sukhatme (1970). The population 1 is stratified in two strata in accordance with

area under wheat in 1936 less than or more than 200 acres. Parameters for this population are: $N = 34$, $N_1 = 20$, $N_2 = 14$, $\delta^* = .7635$, $\delta = .3638$, $\zeta = .3811$, $\theta = .2436$. The Population 2 comprises of production of cereals in South America for the years 1980 (z), 1988 (y_1) and 1989 (y_2), compiled from The Statistical year book, United Nations (1988/89). The population is stratified in two strata considering 1988 production of more or less than 570 (thousand metric tons). The parameters for this population 2 are: $N = 19$, $N_1 = 7$, $N_2 = 12$, $\delta^* = -.6939$, $\delta = .7666$, $\zeta = 1.1478$, $\theta = .3681$. The population 3 compiled by Singh and Chaudhuri (1986) relates to the area under wheat in hectare during 1979-80 (y_2) and 1978-79 (y_1) and total cultivated area in 1978-79 (z) of 16 villages of Meerut District. The parameters for the population 3 are: $N = 16$, $N_1 = 9$, $N_2 = 7$, $\delta^* = .7729$, $\delta = .1057$, $\zeta = .3965$, $\theta = .2827$.

The following table shows relative efficiencies of the proposed Strategies 1, 2 and the one proposed by Prasad and Graham (1994) with respect to Chotai (1974) which are respectively denoted by $E_1 = V_c/M_1$, $E_2 = V_c/M_2$ and $E_3 = V_c/V_{PG}$.

Table 1
Efficiencies of the Strategies

f	Population 1			Population 2			Population 3		
	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3
.05	1.0463	1.1033	1.0181	1.0196	1.0850	.8262	1.0053	1.0864	1.0030
.10	1.0479	1.0895	1.0187	1.0202	1.0711	.8212	1.0055	1.0711	1.0031
.15	1.0496	1.0776	1.0194	1.0209	1.0579	.8172	1.0057	1.0577	.0033
.20	1.0514	1.0683	1.0200	1.0216	1.0519	.8123	1.0058	1.0469	1.0034
.25	1.0533	1.0622	1.0208	1.0224	1.0490	.8071	1.0061	1.0396	1.0035
.30	1.0554	1.0604	1.0216	1.0232	1.0530	.8017	1.0063	1.0368	1.0036

From the above table, we note that in all the three populations, Strategy 2 fares better than the others. It is also worth noting that both the proposed strategies fare better than those of Chotai (1974) and Prasad and Graham (1994). For the population 1, $\zeta = .3811$ which is quite favourable for Prasad and Graham's (1994) strategy, hence for the proposed Strategy 1. Both Prasad and Graham's strategy and Strategy 1, performed better than Chotai's (1974) strategy. For the population 2, $\zeta = 1.1478$ which is high and unfavourable for Prasad and Graham's (1994) strategy, but $\delta = .7666$ is quite favourable to Strategy 1. Hence, for the population 2, Prasad and Graham's strategy becomes less efficient than that of Chotai (1974), but the proposed Strategy 1 remains better. For the population 3, $\zeta = .3965$ which is quite favourable for Prasad and Graham (1994) but at the same time $\delta^* = .7729$ and this (δ^*) favours Chotai (1974). In fact Chotai's (1974) strategy is marginally inferior to Prasad and Graham's (1994) strategy but the proposed Strategy 2 remains better than both. It should be noted that the examples shown here are quite unusual in the

sense that they present low correlation between y_2 and z (in example 1, $\delta = .3638$ and in example 3, $\delta = .1057$) and there is a negative correlation between y_2 and y_1 ($\delta^* = -.6939$) in example 2. The correlations δ and δ^* are expected to be high and positive. Hence, further investigation is needed to compare the performances of the present strategies with suitable data.

Table 2
Sensitivity of Efficiency $E^* = V_{PG}/M_{\tilde{\beta}}$

$ v $.05	.10	.15	.20	.25	.30
Population 1						
0	1.028	1.029	1.030	1.031	1.032	1.033
.2	1.027	1.027	1.028	1.029	1.031	1.032
.4	1.023	1.024	1.027	1.026	1.027	1.028
.6	1.017	1.108	1.019	1.019	1.020	1.021
.8	1.010	1.010	1.010	1.011	1.011	1.011
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.989	.988	.988	.988	.988	.987
1.4	.976	.976	.975	.974	.973	.972
Population 2						
0	1.234	1.241	1.249	1.257	1.266	1.278
.2	1.219	1.227	1.233	1.241	1.249	1.258
.4	1.180	1.186	1.191	1.197	1.204	1.211
.6	1.125	1.128	1.133	1.137	1.141	1.146
.8	1.063	1.065	1.067	1.068	1.070	1.073
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.939	.938	.936	.935	.933	.931
1.4	.883	.880	.877	.875	.871	.869
Population 3						
0	1.002	1.002	1.004	1.003	1.003	1.003
.2	1.002	1.002	1.002	1.002	1.003	1.002
.4	1.002	1.002	1.002	1.002	1.002	1.002
.6	1.001	1.002	1.002	1.002	1.002	1.001
.8	1.001	1.001	1.001	1.001	1.001	1.001
1.0	1.000	1.000	1.000	1.000	1.000	1.000
1.2	.999	.999	.999	.999	.999	.999
1.4	.998	.997	.998	.998	.998	.998

To study the effect of departure of the optimum value of $\beta = \beta_0$ when some guess value of β is used in Strategy 1, one may consider sensitivity of efficiency of \hat{Y}_2 for the Strategy 1 for different choices of β , following Prasad and Srivenkataramana (1980). The minimum variance of \hat{Y}_2 for the Strategy 1 when some guess value of $\beta_0 = \tilde{\beta}$ is used, produces

$$V_{\min}(\hat{Y}_2 | \tilde{\beta}) = k(1 - f + \sqrt{\zeta^{**}}) \sigma_2^2 / 2 = M_{\tilde{\beta}} \quad (9)$$

where $\zeta^{**} = [1 - (1 - v^2) \delta^2] \zeta$ and $v = 1 - \tilde{\beta} / \beta_0$.

From (9), we note that the proposed Strategy 1 with the guess value $\tilde{\beta}$ fares better or worse than Prasad and

Graham's (1994) strategy according to $|v| < 1$ or $|v| > 1$. Similarly, the proposed Strategy 1 with $\beta = \tilde{\beta}$ performs better or worse than Chotai's (1974) strategy according to $v^2 > \text{or} < (1 - 1/\delta^2)(1 - 1/\zeta)$. Table 2 proceeds sensitivity E^* of the estimator \hat{Y}_2 compared to Prasad and Graham's (1994) strategy where $E^* = V_{PG} / M_{\tilde{\beta}}$. From the Table 2, the loss with $v > 1$ is likely to be more than the gain with $v < 1$ for population 1 and population 3 but the situation is reverse for population 2.

CONCLUSION

In sampling over two occasions, one should utilize data collected on the first occasion to get an efficient estimator for the population total on the second occasion. Chotai (1974) used data collected on the first occasion at the stage of estimation, while Prasad and Graham did so at the stage of selection (and hence estimation) of the matched sample. In this article, two strategies have been proposed. The first one utilizes data collected at the first occasion for the selection of the matched sample similar to Prasad and Graham and formation of a regression estimator as determined by Chotai (1974). These make Strategy 1 more efficient than that of Prasad and Graham. The proposed Strategy 2 utilized first occasion values as a stratification variable, measure of size for the selection of the matched sample for the second occasion, and formation of a regression type estimator involving auxiliary variable (z), available on the first occasion. Intuitively one should expect the proposed Strategy 2 to perform better than the others mentioned here, but no theoretical result was established due to the complexity of the expression for the minimum variance of the proposed estimator. However, superiority of the Strategy 2 was established through numerical data.

ACKNOWLEDGEMENTS

The author is grateful to the referee, Associate Editor and the Editor for their valuable comments that substantially improved the earlier version of this paper. This work was supported by the FRD, South Africa.

REFERENCES

- ARNAB, R. (1991). On sampling over two occasions using varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 43(3), 282-290.
- AVADHANI, M.S., and SUKHATME, B.V. (1970). A comparison of two sampling procedures with an application to successive sampling. *Applied Statistics*, 19, 251-259.
- CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā*, Series C, 36, 173-180.

- PRASAD, N.G.N., and SRIVENKATARAMANA, T. (1980). A modification to the Horvitz-Thompson estimator under the Midzuno sampling scheme. *Biometrika*, 67, 709-711.
- PRASAD, N.G.N., and GRAHAM, J.E. (1994). PPS sampling over two occasions. *Survey Methodology*, 20, 59-64.
- RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.
- RAO, J.N.K., HARTLEY, H.O., and COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.
- SINGH, D., and CHAUDHURI, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. India: Wiley Eastern Limited, 166.
- SINGH, M.P. (1967). The relative efficiency of some two-phase sampling schemes. *Annals of Mathematical Statistics*, 38, 937-940.
- SUKHATME, P.V., and SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, 185.
- UNITED NATIONS (1992). *Statistical Year Book*, (1988/89). New York: United Nations, 356.

Confidence Intervals for Proportions With Small Expected Number of Positive Counts Estimated From Survey Data

EDWARD L. KORN and BARRY I. GRAUBARD¹

ABSTRACT

In the nonsurvey setting, "exact" confidence intervals for proportions calculated using the binomial distribution are frequently used instead of intervals based on approximate normality when the number of positive counts is small. With complex survey data, the binomial intervals are not applicable, so intervals based on the assumed approximate normality of the sample-weighted proportion are used, even if the number of positive counts is small. We propose a simple modification of the binomial intervals to be used in this situation. Limited simulations are presented that show the coverage probability of the proposed intervals is superior to that of the normality-based intervals, logit-transform intervals, and intervals based on a Poisson approximation. Applications are given involving the prevalence of Human Immunodeficiency Virus (HIV) based on data from the third National Health and Nutrition Examination Survey, and the proportion of users of cocaine based on data from the Hispanic Health and Nutrition Examination Survey.

KEY WORDS: Binomial confidence interval; Exact confidence interval; Logit transformation; Poisson confidence interval.

1. INTRODUCTION

With complex survey data, the typical construction of a $1 - \alpha$ level confidence interval for a proportion of positive counts for a 0-1 variable is

$$\hat{p} \pm t_d (1 - \alpha/2) [\text{var}(\hat{p})]^{1/2} \quad (1.1)$$

where \hat{p} is the sample-weighted estimator of the proportion, $\text{var}(\hat{p})$ is the variance estimator of \hat{p} , and $t_d(1 - \alpha/2)$ is the $1 - \alpha/2$ quantile of a t distribution with d degrees of freedom. The estimator $\text{var}(\hat{p})$ is computed using linearization or a replication method to reflect the sample design, including the fact that \hat{p} is a sample-weighted estimator. By complex survey data, we mean data obtained from a multistage design with stratified selection of clusters at the first stage. For such a sample design, d is usually taken to be equal to the number of sampled clusters minus the number of strata (Korn and Graubard 1990). The confidence interval (1.1), which we shall refer to as the "linear interval", is based on the assumption that \hat{p} is approximately normally distributed. Under various reasonable asymptotics, this is known to be true (Krewski and Rao 1981). The use of the t quantile rather than a normal-distribution quantile in (1.1) is based on empirical evidence (Frankel 1971, ch. 7), and it can also be formally justified using strong assumptions (Korn and Graubard 1990).

When the expected number of positive counts is small, the approximate normality of \hat{p} breaks down (Cochran 1977, p. 58). For a simple random sample (or in the nonsurvey setting), one can avoid the normality assumption

by using the Clopper and Pearson (1934) confidence interval based on the binomial distribution; see Vollset (1993) for a complete discussion of confidence intervals for proportions in the nonsurvey setting. When x positive responses are seen in a simple random sample of size n , the Clopper-Pearson $1 - \alpha$ level confidence interval $(p_L(x, n), p_U(x, n))$ can be expressed as (Johnson, Kotz and Kemp 1993, p. 130):

$$p_L(x, n) = \frac{v_1 F_{v_1, v_2}(\alpha/2)}{v_2 + v_1 F_{v_1, v_2}(\alpha/2)}$$

$$p_U(x, n) = \frac{v_3 F_{v_3, v_4}(1 - \alpha/2)}{v_4 + v_3 F_{v_3, v_4}(1 - \alpha/2)} \quad (1.2)$$

where $v_1 = 2x, v_2 = 2(n - x + 1), v_3 = 2(x + 1), v_4 = 2(n - x)$ and $F_{d_1, d_2}(\beta)$ is the β quantile of an F distribution with d_1 and d_2 degrees of freedom. For one-sided confidence bounds, α is used instead of $\alpha/2$ in the above expressions. For a simple random sample, these intervals are known to have coverage probability greater than or equal to their nominal level, regardless of the expected number of positive counts. They are sometimes referred to as "exact" confidence intervals; we shall refer to them as the "binomial intervals".

In this paper we suggest a simple modification to the binomial intervals to make them applicable for a proportion estimated from complex survey data. We are especially interested in the situation when the expected number of positive counts is small. Many survey analysts would not

¹ Edward L. Korn, Biometric Research Branch, EPN-739, National Cancer Institute, Bethesda, MD 20892, U.S.A.; Barry I. Graubard, Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.

present estimated proportions in this situation, since they are unreliable. For example, applying the relative-standard-error criterion for presenting proportions in the 1996 National Household Survey on Drug Abuse (SAMHSA 1998), the estimated proportion of women using cocaine in Table 7 would not be presented. We believe such proportions can provide valuable information, but that their lack of precision needs to be explicitly stated by presenting confidence intervals. In section 2, we define our proposed confidence intervals and define intervals based on a logit transformation and the Poisson distribution that have been suggested in the literature. Simulation results are presented in section 3 that compare the intervals. We find that the proposed intervals behave well in terms of coverage probability of the true proportion and in terms of their average width. Two applications are given in section 4 involving large surveys, but where the number of positive counts is expected to be small. We end with a discussion of some related work that constructs confidence intervals that are guaranteed to attain their nominal coverage probability regardless of the population configuration of counts.

2. PROPOSED AND OTHER CONFIDENCE LIMITS

For a $1 - \alpha$ level confidence interval based on a sample of size n , first define the effective sample size by

$$n^* = \frac{\hat{p}(1 - \hat{p})}{\text{var}(\hat{p})} \quad (2.1)$$

and the degrees-of-freedom adjusted effective sample size by

$$n_{df}^* = \frac{\hat{p}(1 - \hat{p})}{\text{var}(\hat{p})} \left(\frac{t_{n-1}(1 - \alpha/2)}{t_d(1 - \alpha/2)} \right)^2 \quad (2.2)$$

Both n^* and n_{df}^* are set equal to n when $\hat{p} = 0$. The proposed limits substitute n_{df}^* for n , and $\hat{p}n_{df}^*$ for x in (1.2), viz. $p_L(\hat{p}n_{df}^*, n_{df}^*)$ and $p_U(\hat{p}n_{df}^*, n_{df}^*)$. (When n is large, the $1 - \alpha/2$ quantile of a normal distribution can be used in place of $t_{n-1}(1 - \alpha/2)$ in (2.2).) For estimating a confidence interval for a proportion on a subdomain of the population, the sample size n is taken to be equal to the sample size restricted to the subdomain.

A heuristic justification for this procedure is as follows. The effective sample size (2.1) is n divided by an estimator of the design effect of the survey. This seems to be a reasonable way to incorporate the additional variability of \hat{p} due to the complex sampling. For confidence interval construction, the variability of the variance estimator is also important. The second fraction in (2.2) takes into account the fact that $\text{var}(\hat{p})$ will typically be more variable than a variance estimator that would be used for simple random sampling. If d is large, then this factor is close to one and unneeded. For small d and large n and $\hat{p}n_{df}^*$, we would like

the proposed interval to be close to the interval (1.1), which is appropriate in this situation. Using the fact that $F_{u,w}(\beta) \approx 1 + z(\beta) \sqrt{2(1/u + 1/w)}$ for large u and w (Johnson and Kotz 1970, p. 81), this is true, i.e., $\hat{p} - p_L(\hat{p}n_{df}^*, n_{df}^*) \approx p_U(\hat{p}n_{df}^*, n_{df}^*) - \hat{p} \approx t_d(1 - \alpha/2)[\text{var}(\hat{p})]^{1/2}$.

A procedure closely related to the proposed procedure was developed by Breeze (1990) for use in the U.K. General Household Survey. This procedure is based on the simple-random-sampling $1 - \alpha$ confidence interval $(po_L(x), po_U(x))$ for a Poisson random variable x , which can be expressed as (Johnson *et al.* 1993, p. 171):

$$po_L(x) = 0.5 \chi_{v_1}^2(\alpha/2) \text{ and } po_U(x) = 0.5 \chi_{v_2}^2(1 - \alpha/2)$$

where $v_1 = 2x$, $v_2 = 2(x + 1)$, and $\chi_v^2(\beta)$ is the β quantile of a χ^2 distribution with v degrees of freedom. With complex survey data, the confidence interval is taken to be $(po_L(\hat{p}n^*)/n^*, po_U(\hat{p}n^*)/n^*)$.

A third procedure for confidence interval construction is based on a logit transform. For a $1 - \alpha$ level confidence interval, the interval is

$$\left(\frac{1}{1 + \exp(-LLOGIT)}, \frac{1}{1 + \exp(-ULOGIT)} \right)$$

where

$$LLOGIT = \log \frac{\hat{p}}{1 - \hat{p}} - t_d(1 - \alpha/2) \frac{[\text{var}(\hat{p})]^{1/2}}{\hat{p}(1 - \hat{p})} \quad (2.3)$$

and

$$ULOGIT = \log \frac{\hat{p}}{1 - \hat{p}} + t_d(1 - \alpha/2) \frac{[\text{var}(\hat{p})]^{1/2}}{\hat{p}(1 - \hat{p})} \quad (2.4)$$

These intervals, with a normal-distribution quantile instead of a t distribution quantile, were suggested for use with the 1996 National Household Survey on Drug Abuse (SAMHSA 1998). When $\hat{p} = 0$, in the nonsurvey setting one might add a small constant to the observed number of events and nonevents, e.g., $1/2$, to be able to calculate the logit-transform confidence interval (Agresti 1990, pp. 249-250). In the present setting, when $\hat{p} = 0$, we set the confidence interval equal to the binomial interval $(p_L(0, n), p_U(0, n))$.

In applications where it is known before sampling that the (true) design effect will be greater than 1, various modifications of the above procedures are possible. For our proposal, we recommend in this situation truncating the degrees-of-freedom adjusted effective sample size at n . That is, if n_{df}^* is greater than n , we set its value to n , and define the lower and upper confidence limits to be $p_L(\hat{p}n, n)$ and $p_U(\hat{p}n, n)$. For the Breeze intervals, one could set n^* to be n if $n^* > n$. For the linear or logit intervals, one can use the simple-random-sampling variance estimator $\hat{p}(1 - \hat{p})/n$ in place of $\text{var}(\hat{p})$ in (1.1), (2.3) and (2.4) if $n^* > n$; see SAMHSA (1998) for additional truncation suggestions. The justification of these truncation

procedures is that the design effect may be estimated to be less than one because of instability of the variance estimator $\hat{\text{var}}(\hat{p})$. This type of instability may be especially large because \hat{p} is small (SAMHSA 1998). The effect of these truncation procedures is to make the confidence intervals wider and more conservative. In theory, one could also adjust the estimated effective sample sizes when it is known before the sampling that the (true) design effect is less than one. However, to be conservative, we do not recommend doing this.

Our focus in this paper is on confidence intervals for the "superpopulation" probability that the outcome $Y = 1$ rather than the finite-population proportion. That is, the target parameter is $p = \sum_{u=1}^N p_u / N$ rather than $P = \sum_{u=1}^N Y_u / N$, where Y_u has a Bernoulli distribution with parameter p_u , and N is the population size. The simulated coverage probabilities given in the next section therefore refer to coverage of p . With this target parameter in mind, we do not use finite-population correction factors when estimating $\hat{\text{var}}(\hat{p})$ for use in (2.2); additional adjustments to the design-based variance $\hat{\text{var}}(\hat{p})$ for superpopulation inference are not pursued here (Korn and Graubard 1998). A referee suggests the possibility of a model-based approach to estimating a confidence interval for p . However, in our limited experience, such approaches yield estimators similar to weighted estimators and offer no advantages for

inference (Pfeffermann and LaVange 1989; Graubard and Korn 1996).

If one were interested in a confidence interval for P , we would recommend using the proposed intervals but with $\hat{\text{var}}(\hat{p})$ in (2.2) containing the finite-population correction factors. A confidence interval for $\sum_{u=1}^N Y_u$ could be obtained by multiplying the ends of the confidence interval for P by N , if known, or by an estimator \hat{N} of N , if not known. (In theory, one could account for the variability of \hat{N} , but this additional variability will be small.) An alternative approach for estimating a confidence interval for P would be to modify the usual limits (Guenther 1983) appropriate for a simple random sample (based on the hypergeometric distribution) similarly to the way the proposed intervals modify the binomial intervals.

3. SIMULATIONS

The main simulation results are presented in Tables 1-5. Table 1 presents the results of simulations in which datasets of 32 clusters, each with sample size 100, were simulated. Within cluster i , the number of positive events was simulated with a binomial distribution with probability parameter p_i . In Table 1, we refer to the $\{p_i, i = 1, \dots, 32\}$ as the cluster probabilities. For the top third of the table, the cluster probabilities are taken to be the constant $p = .1, .02$,

Table 1
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Sample Weights are 1 Or 10 with Probability 1/2 (Noninformative Sample Weights)

Distribution of cluster proportions ^a	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1	.1	320	4.6	5.5	5.3	4.6	4.5	4.1	4.8	4.4
.02	.02	64	3.4	7.1	5.2	4.6	4.5	4.7	4.2	4.4
.01	.01	32	2.9	8.0	5.4	4.5	4.4	4.5	4.0	4.1
.0025	.0025	8	1.6	9.5	5.5	1.8	3.6	2.2	3.3	1.8
(1/2, 1/2)										
.05, .15	.1	320	4.3	5.8	5.5	4.3	4.3	3.8	4.7	4.1
.01, .03	.02	64	3.1	7.5	5.2	4.8	4.3	4.8	4.0	4.5
.005, .015	.01	32	2.7	8.6	5.2	4.7	4.1	4.9	3.7	4.4
.00125, .00375	.0025	8	1.5	9.9	5.4	2.0	3.4	2.3	3.1	2.0
(3/4, 1/4)										
.05, .25	.1	320	3.1	7.8	4.7	5.6	3.4	5.0	3.6	5.3
.01, .05	.02	64	2.7	8.6	5.1	5.3	4.0	5.4	3.7	5.0
.005, .025	.01	32	2.2	9.8	5.0	5.3	3.7	5.5	3.3	5.0
.00125, .00625	.0025	8	1.3	10.7	5.3	2.2	3.3	2.5	3.0	2.2

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

Table 2
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Informative Sample Weights are 1 or 10 (See Text)

Distribution of cluster proportions ^a	Overall weighted proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1	.1	191.0	4.3	5.9	5.1	4.9	4.2	4.4	4.6	4.6
.02	.02	36.9	3.3	7.3	5.3	4.3	4.4	4.4	4.1	4.1
.01	.01	18.4	2.8	8.7	5.5	4.0	4.3	4.3	3.9	3.7
.0025	.0025	4.6	1.3	18.7	6.1	4.8	3.2	4.8	2.8	4.8
(1/2, 1/2)										
.05, .15	.1	191.0	5.0	5.0	6.4	3.7	5.1	3.2	5.4	3.4
.01, .03	.02	36.9	3.0	7.9	5.4	4.5	4.3	4.6	4.0	4.3
.005, .015	.01	18.4	2.5	9.2	5.4	4.2	4.1	4.4	3.7	3.9
.00125, .00375	.0025	4.6	1.3	19.0	6.1	4.9	3.2	4.9	2.8	4.9
(3/4, 1/4)										
.05, .25	.1	191.0	4.7	5.7	7.1	4.1	5.1	3.6	5.5	3.8
.01, .05	.02	36.9	2.6	8.9	5.2	5.2	4.0	5.3	3.7	4.9
.005, .025	.01	18.4	2.1	10.1	5.3	4.8	3.8	5.1	3.4	4.5
.00125, .00625	.0025	4.6	1.2	19.8	5.9	5.3	3.2	5.3	2.8	5.3

(a) Fractions in parentheses are the probabilities that the cluster weighted proportions have the stated value.

Table 3
Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32 Clusters and 100 Observations Per Cluster; Unweighted Analyses

Distribution of cluster proportions ^a	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1	.1	320	5.0	4.9	5.7	4.2	4.9	3.8	5.2	4.1
.02	.02	64	3.8	6.3	5.2	4.5	4.7	4.8	4.4	4.4
.01	.01	32	3.5	6.8	5.6	4.4	4.7	4.4	4.3	4.0
.0025	.0025	8	2.5	8.8	5.6	3.8	4.1	3.9	3.9	3.9
(1/2, 1/2)										
.05, .15	.1	320	4.5	5.6	5.6	4.2	4.5	3.7	4.8	4.0
.01, .03	.02	64	3.4	7.0	5.1	4.8	4.5	4.9	4.1	4.6
.005, .015	.01	32	3.0	7.6	5.2	4.8	4.4	4.8	3.9	4.4
.00125, .00375	.0025	8	2.2	9.2	5.4	4.3	3.8	4.3	3.5	4.3
(3/4, 1/4)										
.05, .25	.1	320	3.3	7.7	4.8	5.6	3.5	5.1	3.7	5.3
.01, .05	.02	64	2.9	8.1	5.1	5.2	4.1	5.3	3.8	4.9
.005, .025	.01	32	2.5	9.2	4.9	5.6	3.9	5.6	3.5	5.2
.00125, .00625	.0025	8	2.0	10.4	5.3	5.1	3.8	5.1	3.3	5.1

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

Table 4
 Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32
 Clusters and 10 Observations Per Cluster; Sample Weights are 1 or 10 with Probability 1/2
 (Noninformative Sample Weights)

Distribution of cluster proportions ^a	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.2	.2	64	4.0	6.6	5.2	4.7	3.1	4.7	4.2	4.3
.1	.1	32	3.2	7.8	5.3	4.4	3.6	3.8	3.9	4.0
.025	.025	8	1.7	10.2	5.5	2.1	3.4	2.1	3.2	2.4
(1/2, 1/2)										
.1, .3	.2	64	3.6	7.0	5.0	4.9	2.8	3.4	3.9	4.4
.05, .15	.1	32	3.0	8.1	5.1	4.6	3.4	4.0	3.7	4.2
.0125, .0375	.025	8	1.6	10.6	5.4	2.1	3.3	2.1	3.1	2.5
(3/4, 1/4)										
.1, .5	.2	64	3.1	7.8	4.6	5.3	2.4	3.9	3.3	4.8
.05, .25	.1	32	2.5	9.2	4.8	5.2	3.0	4.6	3.3	4.8
.0125, .0625	.025	8	1.5	11.5	5.3	2.4	3.2	3.5	3.0	2.8

(a) Fractions in parentheses are the probabilities that the cluster proportions have the stated value.

Table 5
 Simulated Lack of Coverage (Percent) of Upper and Lower One-sided 95% Confidence Bounds for Sample Design of 32
 Clusters and 10 or 100 Observations Per Cluster with Probability 1/2; Sample Weights are 1 or 10 with Probability 1/2
 (Noninformative Sample Weights)

Distribution of cluster proportions ^a	Overall proportion	Expected number positive	Method of calculating confidence bounds							
			Linear		Logit		Breeze		Proposed	
			Lower	Upper	Lower	Upper	Lower	Upper	Lower	Upper
(1)										
.1818	.1818	320	5.1	6.0	5.7	5.2	4.2	4.1	5.2	5.0
.0364	.0364	64	4.1	7.6	5.7	5.2	5.0	5.2	4.8	4.9
.0182	.0182	32	3.4	8.5	5.7	5.0	4.7	5.1	4.4	4.7
.0045	.0045	8	2.0	12.7	5.9	3.4	4.0	4.3	3.6	3.8
(1/2, 1/2)										
.0909, .2727	.1818	320	5.0	6.4	6.1	4.8	4.2	3.6	5.2	4.4
.0182, .0545	.0364	64	3.9	8.1	6.0	5.1	4.9	5.0	4.7	4.8
.0091, .0273	.0182	32	3.1	9.3	5.8	5.2	4.5	5.3	4.2	4.9
.0023, .0068	.0045	8	1.8	13.2	5.9	3.6	3.9	4.5	3.5	4.0
(3/4, 1/4)										
.0909, .4545	.1818	320	3.1	9.9	4.6	7.6	2.5	6.3	3.3	7.1
.0182, .0909	.0364	64	2.8	10.9	5.3	7.3	3.9	7.3	3.7	7.0
.0091, .0455	.0182	32	2.4	11.5	5.4	6.8	3.9	6.9	3.6	6.5
.0023, .0114	.0045	8	1.6	14.5	5.7	4.0	3.7	5.0	3.3	4.4

(a) Fractions in parentheses are the probabilities that the cluster weighted proportions have the stated value.

.01, or .0025, corresponding to an expected number of positive events equal to 320, 64, 32, or 8 out of the sample size of 3200. For the middle third of the table, the cluster probabilities are taken to be $p/2$ with probability $1/2$ or $3p/2$ with probability $1/2$, with p as in the first third of the table. Varying the p_i across the clusters induces an intraclass correlation among the observations. For the middle third of the table, these correlations (ignoring the sample weights) are .00278, .0051, .0025 and .0006 corresponding to the expected number of positive events being 320, 64, 32, or 8, respectively. For the bottom third of the table, the cluster probabilities are taken to be $p/2$ with probability $3/4$ or $5p/2$ with probability $1/4$, corresponding to intraclass correlations of .0833, .0153, .0076 and .0019. For all simulations in Table 1, sample weights of 1 or 10 are randomly assigned with probability $1/2$ to each observation (noninformative weights).

The results presented in Table 1 are appropriate for one-sided 95% upper and lower confidence limits; ideally the lack-of-coverage percentages in the table should be less than or equal to the nominal value of 5.0. The results are also relevant for two-sided 90% confidence intervals, for which ideally both the upper and lower values in the table should both be ≤ 5.0 (Jennings 1987). For each line of the table, 100,000 datasets were simulated using the random number generator in SAS (1990, p. 631) to estimate the probabilities of noncoverage of the confidence limits.

For the linear confidence bounds, the upper confidence limit falls below the true value more than the 5% nominal level. Somewhat surprisingly, this is true even with as large as an expected 320 positive counts, especially with positive intraclass correlation (middle and bottom third of the table). For the logit-transform confidence bounds, the noncoverage appears slightly higher than the nominal level, especially for the lower limits. Both the Breeze and proposed confidence bounds appear generally conservative. Simple-random-sampling binomial limits are not appropriate for the cases simulated in Table 1 because of the sample weights and the intraclass correlation (in the bottom two-thirds of the table). This can be demonstrated by noting that the lack of coverage for both the upper and lower binomial bounds are greater than 8% for all the cases considered in the table (results not shown).

As it is slightly complicated to discuss confidence interval "lengths" for one-sided bounds, we restrict discussion to the lengths of the two-sided 90% confidence intervals. Over all the simulations presented in Table 1, the Breeze and proposed intervals are 3.3% and 4.9% wider on average than the logit-transform intervals.

Table 2 presents simulation results for the same setup as Table 1 except that the sample weights were taken to be informative. This was done by setting the sample weight to be 10 with probability $2/3$ if the event was positive and with probability $1/3$ if the event was not positive, otherwise the weight was set to 1. The probability that an event was positive in each cluster was adjusted downwards so that the overall

weighted proportions were the same as in Table 1. The results in Table 2 look similar to those in Table 1 except the linear and logit intervals tend to have worse coverage probabilities.

Table 3 presents simulation results for the same setup as Table 1 except the analysis is unweighted. The results are very similar to the Table 1 results. Since the top third of Table 3 corresponds to no intraclass correlation, one could also use the simple-random-sampling binomial limits there. Averaging over the four situations in this third of the table, the proposed limits are 2.5% wider than the binomial limits (results not shown). As the true design effect is 1.0 in the top third of Table 3, these simulations can be used to examine the effect of truncation of n_{df}^* in the proposed procedure. (Truncation is uncommon in the simulations in Table 1, since the true design effects there are all >1 .) Simulation using the proposed procedure with truncation lead to wider more conservative intervals than for the proposed intervals in the top third of Table 3. Averaging over the four situations considered, the proposed limits with truncation are 4.0% wider than the proposed limits (results not shown for truncated limits).

Table 4 presents simulation results for the same setup as Table 1 except 10 rather than 100 observations are simulated within each cluster. The results are very similar to Table 1 when one compares simulations with the same expected number of positive events. The one exception is the increased conservativeness of the Breeze intervals as compared to the proposed method. This is because the overall proportions are higher in Table 4 than Table 1 for a given expected number of positive events (since the sample size is smaller in Table 4). The Poisson intervals of Breeze do not work well with proportions that are not small. For example, we performed a simulation corresponding to the top third of Table 1 except that the overall proportion $p = .5$ with 1600 expected number of positive events. The simulated lower and upper lack-of-coverage percentages for the Breeze bounds were 1.2% and 1.3%, compared to 4.6% and 4.7% for the proposed method. The Breeze intervals were on average 37% wider than the proposed intervals.

The Breeze intervals also do not work well when the number of clusters is very small, since they do not account for degrees of freedom of the variance estimation. For example, we performed a simulation corresponding to the top third of Table 1 except that data from only 8 clusters were simulated (with 100 observations per cluster), and $p_i = .1$ so that the expected number of positive events was 80. The simulated lower and upper lack-of-coverage percentages for the Breeze bounds were 6.1% and 5.4%, compared to 4.7% and 4.0% for the proposed method.

Table 5 presents simulation results for the same setup as Table 1 except the cluster sizes were taken to be 10 or 100 with probability $1/2$. The lack-of-coverage probabilities are larger than the nominal 5% in the bottom third of the table for all the methods. The logit intervals also do not behave as well as in Table 1 for the top two-thirds of the table.

An additional set of simulations was done in which two clusters (each of sample size 50) were simulated from each 32 strata. The expected numbers of positive event were taken as in Table 1, the weights were randomly set to 1 or 10, and the probability of a positive event was taken to be different in the different strata to simulate an intracluster correlation. The results (not shown) were very similar to the results given in Table 1.

4. APPLICATIONS

In this section we consider two applications in which the numbers of positive counts are small. In the first application, involving estimating HIV positivity in an unselected population, the numbers of positive counts are small because the rates of HIV infection are small. In the second application, involving estimating whether individuals have ever used cocaine, the rates are not small but the numbers of positive counts are small because we restrict the analyses to relatively small subdomains. For both applications, SUDAAN (Shah, Barnwell and Bieler 1995) was used to calculate the (design-based) standard errors of the proportions, and the function "FINV" in SAS (1990, p. 547) was used to calculate the quantiles of the F distribution in (1.2).

4.1 Seroprevalence of HIV Estimated From the Third National Health and Nutrition Examination Survey (NHANES III)

NHANES III was a survey conducted in 1988-1994 of the civilian noninstitutionalized population ages 2 months or older of the United States (National Center for Health

Statistics 1994). An HIV test was performed on participating individuals 18 years of age or older. McQuillan, Khare, Karon, Schable and Vlahov (1997) studied the seroprevalence of HIV for individuals under the age of 60 years and various subgroups, some of which are displayed in Table 6. Of the 11,202 individuals tested, 59 were infected. The estimated prevalence in Table 6, 0.32%, is far from the unweighted proportion, $0.53\% = 59/11202$, because the estimated prevalence is a weighted proportion utilizing the sample weights. Because the testing for HIV was anonymous, for these analyses the sample weights were derived from the original NHANES III sample weights of all individuals in the same stand (survey location), race/ethnicity group, sex, and age group (18-39 vs. 40-59) of the tested individual (M. Khare, personal communication). The pseudo-design for variance estimation was the sampling of 2 pseudo-PSU's from each of 23 strata (M. Khare, personal communication), which is not the pseudo-design typically used for NHANES III variance estimation.

The linear 90% confidence intervals for prevalence for the various groups in Table 6 are shifted to the left and shorter than the other intervals, which are similar to each other. The proposed intervals are very slightly wider than the Breeze or logit intervals. The effective sample sizes calculated in (2.1) are markedly smaller than the sample sizes because of the design effects of the survey; the confidence intervals based on the truncated procedures will therefore be identical to the ones given in Table 6. The differences between n^* and n_{df}^* are relatively minor. For this relatively rare outcome, the simulations given in section 3 suggest that the Breeze and proposed confidence intervals may maintain their nominal 90% coverage probabilities better than the other intervals.

Table 6
Seroprevalence of HIV Among Adults Aged 18-59 Years Based on the Third National Health and Nutrition Examination Survey

	Total	Sex		Race/ethnicity		
		Male	Female	White	Black	Mex. - Amer.
Sample size	11202	5142	6060	4128	3579	3495
Number infected	59	44	15	9	38	12
Prevalence (%) \pm SE	0.320 \pm 0.076	0.519 \pm 0.130	0.127 \pm 0.053	0.203 \pm 0.071	1.100 \pm 0.247	0.368 \pm 0.134
Effective sample size						
n^*	5588	3056	4433	3976	1779	2039
n_{df}^*	5148	2816	4084	3664	1640	1880
Linear 90% con. int.	(0.19, 0.45)	(0.30, 0.74)	(0.04, 0.22)	(0.08, 0.33)	(0.68, 1.52)	(0.14, 0.60)
Logit 90% con. int.	(0.21, 0.48)	(0.34, 0.80)	(0.06, 0.26)	(0.11, 0.37)	(0.75, 1.62)	(0.20, 0.69)
Breeze 90% con. int.	(0.21, 0.48)	(0.32, 0.79)	(0.05, 0.26)	(0.10, 0.37)	(0.73, 1.61)	(0.18, 0.68)
Proposed 90% con. int.	(0.20, 0.48)	(0.32, 0.80)	(0.05, 0.26)	(0.10, 0.37)	(0.71, 1.63)	(0.17, 0.69)

Table 7
 "Ever Users" of Cocaine Among Adults Ages 12-44 Years Based on Individuals with 16 or More Years of Education
 Sampled in Hispanic Health and Nutrition Examination Survey

	Total	Sex	
		Male	Female
Sample size	123	69	54
Ever-users	13	10	3
Proportion (%) \pm SE	11.6 \pm 2.5	14.3 \pm 3.4	7.0 \pm 4.8
Effective sample size			
n^*	167.1	105.0	28.2
n_{df}^*	132.8	84.4	22.9
Linear 90% confidence int.	(7.0, 16.2)	(8.0, 20.7)	(-1.9 ^a , 15.9)
Logit 90% confidence int.	(7.8, 17.1)	(9.1, 21.9)	(1.9, 22.8)
Breeze 90% confidence int.	(7.7, 17.0)	(8.3, 23.2)	(0.9, 24.8)
Proposed 90% confidence int.	(7.4, 17.2)	(8.5, 22.1)	(0.9, 22.7)
Truncated Procedures			
Linear 90% confidence int.	(6.3, 17.0)	(6.5, 22.2)	same as above
Logit 90% confidence int.	(7.2, 18.2)	(8.1, 24.1)	"
Breeze 90% confidence int.	(7.1, 18.1)	(7.7, 24.4)	"
Proposed 90% confidence int.	(7.2, 17.5)	(8.0, 23.2)	"

(a) In practice, this interval would be presented as (0, 15.9) since negative proportions are impossible.

4.2 Use of Cocaine Among College-educated Individuals Sampled in the Hispanic Health and Nutrition Examination Survey (HHANES)

HHANES was a survey conducted in 1982-1983 of three Hispanic groups living in the United States (National Center for Health Statistics 1985). We restrict attention here to the Mexican-American sample. Individuals ages 12-44 years were asked "About how old were you the first time you tried cocaine?". The possible answers were the age of the individual (in years) when he first tried cocaine, a "never used" category, and a "don't know" category. We consider estimating the proportion of "ever-users" among individuals who completed 16 or more years of education (for which there were no "don't know" responses).

There were 13 ever-users among 123 sampled individuals, with the sample-weighted proportion being 11.6% (Table 7). The design-based standard error, 2.5%, is estimated with only 8 degrees of freedom since the sampling design of HHANES can be approximated by the sampling of 2 PSU's from each of 8 strata (Kovar and Johnson 1986). The effective sample sizes are $n^* = 167.1$ and $n_{df}^* = 132.8$, which are both greater than the sample size. This is because the estimated design effect is .736, so that $n^* = 123/.736 = 167.1$. (The second factor in (2.2) is 0.794.) Despite the stratification, we think that the true design effect is greater than 1 for this survey because of the clustering and the sample weighting. (The estimated design effect is estimated poorly because of the limited degrees of freedom.) We therefore think that the truncated procedures are reasonable for this application.

Because of the limited degrees of freedom, and because the outcome is not rare, there are more differences between the logit, Breeze and proposed confidence intervals displayed in Table 7. Based on the simulations given in section 3, we recommend the proposed (truncated) confidence intervals.

Our approach may appear slightly inconsistent for this survey in that we accept poorly-estimated effective sample sizes less than the sample size but truncate those greater. We believe that this is a reasonable conservative approach to use when it is thought that the true design effect is probably greater than 1.

5. DISCUSSION

Although the confidence intervals proposed here had adequate coverage probability for almost all the simulations performed, this is not guaranteed for all possible configurations of the population, *e.g.*, see the bottom third of Table 5. An example with a more serious lack of coverage can also easily be constructed: Suppose that the population consists of clusters of size 100, and that 10% of the clusters have all positive units and the remaining 90% have all zero units. If we sample 10 clusters as a simple random sample, and subsample all the units in the sampled clusters, then 35% ($= (1-.1)^{10}$) of the time we will observe no positive units in the sample size of 1000. In this situation, our proposed intervals reduce to the usual binomial ones, so that, *e.g.*, the upper 95% confidence limit for the population proportion is given by .003 ($= 1-.05^{1/1000}$). This implies that

the upper 95% confidence interval is less than the true value of .10 at least 35% of the time, a serious undercoverage.

It is possible in simple sampling situations to construct confidence intervals that are guaranteed to have at least their nominal coverage probability by considering all possible configurations of the population, and using a least-favorable configuration for the coverage probability. For the hypothetical single-stage cluster sample mentioned above, for example, an upper 95% confidence limit could be given by the binomial limit based on 0 positive units out of 10, *i.e.*, .26 ($=1-.05^{1/10}$). Such confidence intervals, which can become computationally intensive to calculate, have been studied by Gross and Frankel (1991), who also suggest some less computationally intensive approximations.

The advantages of our proposed intervals over such approaches are (1) they are easy to calculate, (2) they accommodate any complex sampling design, including nonresponse and poststratification adjustments to the sample weights, (3) they will generally maintain their nominal coverage probability, (4) they will be less conservative than intervals that are guaranteed to maintain their nominal coverage probability for all population configurations, and (5) they have better properties than the linear intervals, logit-transform or Breeze intervals. Conclusions (2) and (5) are based on our simulation results, which of course do not cover all possible situations. More research would be useful in this regard.

ACKNOWLEDGEMENTS

The authors like to thank M. Khare for providing the prevalence estimates and their design-based standard errors given in Table 6, and the Associate Editor and referees for their helpful comments.

REFERENCES

- AGRESTI, A. (1990). *Categorical Data Analysis*. New York: Wiley.
- BREEZE, E. (1990). *General Household Survey: Report on Sampling Error*. London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).
- CLOPPER, C.J., and PEARSON, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404-413.
- COCHRAN, W.G. (1977). *Sampling Techniques*. Third Edition. New York: Wiley.
- FRANKEL, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation*. Ann Arbor, MI: Institute for Social Research, The University of Michigan.
- GRAUBARD, B.I., and KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5, 263-281.
- GROSS, S.T., and FRANKEL, M.R. (1991). Confidence limits for small proportions in complex samples. *Communications in Statistics – Theory and Methods*, 20, 951-975.
- GUENTHER, W.C. (1983). Hypergeometric distributions. In *Encyclopedia of Statistical Sciences*, Volume 3, (Eds. S. Kotz and N.L. Johnson). New York: Wiley, 707-712.
- JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *American Statistician*, 41, 335-337.
- JOHNSON, N.L., and KOTZ, S. (1970). *Continuous Univariate Distributions – 2*. New York: Wiley.
- JOHNSON, N.L., KOTZ, S., and KEMP, A.W. (1993). *Univariate Discrete Distributions*. Second Edition. New York: Wiley.
- KORN, E.L., and GRAUBARD, B.I. (1990). Simultaneous testing or regression coefficients with complex survey data: use of Bonferroni *t*-statistics. *American Statistician*, 44, 270-276.
- KORN, E.L., and GRAUBARD, B.I. (1998). Variance estimation for superpopulation parameters. *Statistica Sinica*, 8, 1131-1151.
- KOVAR, M.G., and JOHNSON, C. (1986). Design effects from the Mexican American portion of the Hispanic Health and Nutrition Examination Survey: a strategy for analysts. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-399.
- KREWSKI, D., and RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.
- McQUILLAN, G.M., KHARE, M., KARON, J.M., SCHABLE, C.A., and VLAHOV, D. (1997). Update on the seroepidemiology of Human Immunodeficiency Virus in the United States household population: NHANES III, 1988-94. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 14, 355-360.
- NATIONAL CENTER FOR HEALTH STATISTICS (1985). Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. *Vital and Health Statistics* 1(19). Hyattsville, MD: National Center for Health Statistics.
- NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics* 1(32). Hyattsville, MD: National Center for Health Statistics.
- PFEFFERMANN, D., and LAVANGE, L. (1989). Regression models for stratified multi-stage cluster samples. In *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, and T.M.F. Smith). New York: Wiley, 237-260.
- SAMHSA (1998). National Household Survey on Drug Abuse: Main Findings 1996. (DHHS Publication No. (SMA) 98-3200). Rockville, MD: SAMHSA.
- SAS (1990). *SAS Language: Reference, Version 6*, First Edition. Cary, NC: SAS Institute Inc.
- SHAH, B.V., BARNWELL, B.G., and BIELER, G.S. (1995). *SUDAAN User's Manual, Release 6.40*. Research Triangle Park, NC: Research Triangle Institute.
- VOLLSET, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12, 809-824.

ACKNOWLEDGEMENTS

Survey Methodology wishes to thank the following persons who have served as referees during 1998. An asterisk indicates that the person served more than once.

- B.M. Balk, *Statistics Netherlands*
- * D.R. Bellhouse, *University of Western Ontario*
- J. Bethel, *Westat, Inc.*
- P. Biemer, *Research Triangle Institute*
- * D.A. Binder, *Statistics Canada*
- J.-R. Boudreau, *Statistics Canada*
- K. Brewer, *Australian National University*
- J.M. Brick, *Westat, Inc.*
- A. Brinkley, *U.S. Bureau of the Census*
- T.W. Broene, *Energy Information Administration, U.S.A.*
- N. Buck, *University of Essex*
- P. Cantwell, *U.S. Bureau of the Census*
- * R. Chambers, *University of Southampton*
- * M. Cohen, *National Center for Education Statistics*
- B. Cox, *Mathematica Policy Research*
- * J. Denis, *Statistics Canada*
- * J.-C. Deville, *Institut national de la statistique et des études économiques*
- * P. Dick, *Statistics Canada*
- A. Dorfman, *U.S. Bureau of Labor Statistics*
- * J.D. Drew, *Statistics Canada*
- J. Dufour, *Statistics Canada*
- * J. Eltinge, *Texas A&M University*
- R. Evans, *University of Nebraska - Lincoln*
- M. Feder, *University of Southampton*
- * W.A. Fuller, *Iowa State University*
- * J. Gambino, *Statistics Canada*
- B. Graubard, *U.S. National Cancer Institute*
- * R.M. Groves, *University of Maryland*
- S.J. Haslett, *Massey University, New Zealand*
- L. Hattersley, *Office for National Statistics, U.K.*
- * M.A. Hidioglou, *Statistics Canada*
- * D. Holt, *Office for National Statistics, U.K.*
- K. Humphreys, *University of Glasgow*
- J.-S. Hwang, *Academia Sinica*
- W. Jocelyn, *Statistics Canada*
- E. Johnson, *Educational Testing Service, U.S.A.*
- D. Judkins, *Westat, Inc.*
- * G. Kalton, *Westat, Inc.*
- B. King, *Case Western Reserve University*
- * P.S. Kott, *National Agricultural Statistics Service*
- M. Kovačević, *Statistics Canada*
- M. Kramer, *U.S. Bureau of the Census*
- A. Krieger, *University of Pennsylvania*
- * R. Lachapelle, *Statistics Canada*
- * P. Lahiri, *University of Nebraska - Lincoln*
- M. Larson, *Harvard University*
- * P. Lavallée, *Statistics Canada*
- J. Lawless, *University of Waterloo*
- G. Lee, *Australian Bureau of Statistics*
- * S. Linacre, *Australian Bureau of Statistics*
- * R. Little, *University of Michigan*
- * S. Lohr, *Arizona State University*
- * H. Mantel, *Statistics Canada*
- P.L. do Nascimento Silva, *IBGE, Brasil*
- * G. Nathan, *Central Bureau of Statistics, Israel*
- D. Paton, *Statistics Canada*
- * D. Pfeffermann, *Hebrew University*
- N.G.N. Prasad, *University of Alberta*
- * B. Quenneville, *Statistics Canada*
- T.E. Raghunathan, *University of Michigan*
- * J.N.K. Rao, *Carleton University*
- J.O. Ramsey, *McGill University*
- E. Rancourt, *Statistics Canada*
- * L.-P. Rivest, *Université Laval*
- G. Roberts, *Statistics Canada*
- G. Robinson, *Commonwealth Scientific and Industrial Research Organisation*
- * I. Sande, *Bell Communications Research, U.S.A.*
- C.-E. Särndal, *Université de Montréal*
- * N. Schenker, *University of California - Los Angeles*
- * F.J. Scheuren, *George Washington University*
- A.J. Scott, *University of Auckland*
- * J. Sedransk, *Case Western Reserve University*
- * M.P. Singh, *Statistics Canada*
- * R. Sitter, *Simon Fraser University*
- * C.J. Skinner, *University of Southampton*
- K.P. Srinath, *ABT Associates, Inc.*
- L. Stokes, *University of Texas - Austin*
- * D. Stukel, *Statistics Canada*
- A. Théberge, *Statistics Canada*
- S. Thivierge, *Statistics Canada*
- R. Thomas, *Carleton University*
- I. Thomsen, *Statistics Norway*
- R. Tortora, *Gallup Organization*
- C. Tucker, *U.S. Bureau of Labor Statistics*
- * R. Valliant, *U.S. Bureau of Labor Statistics*
- * V.K. Verma, *University of Essex*
- * P.J. Waite, *U.S. Bureau of the Census*
- * J. Waksberg, *Westat, Inc.*
- * K.M. Wolter, *National Opinion Research Center*
- F. Yu, *Australian Bureau of Statistics*
- E. Zanutto, *University of Pennsylvania*
- * A. Zaslavsky, *Harvard University*

Acknowledgements are also due to those who assisted during the production of the 1998 issues: J. Beauseigle (Dissemination Division) and L. Perreault (Official Languages and Translation Division). Finally we wish to acknowledge D. Blair, C. Larabie, D. Lemire, C. Marleau and G. Ray of Household Survey Methods Division, for their support with coordination, typing and copy editing.

CONTENTS

TABLE DES MATIÈRES

Volume 26, No. 3, September/septembre 1998

Geoffrey WATSON

On the role of statistics in the paleomagnetic proof of continental drift

L. BILLARD

Some statistical opportunities in the agricultural sciences

Noel CRESSIE

Transect-spacing design of ice cores on the Antarctic continent

Daniel GERVINI and Victor J. YOHAI

Robust estimation of variance components

David R. BRILLINGER and Brent S. STEWART

Elephant seal movements: modelling migration

Phaik Mooi LEONG and Stephan MORGENTHALER

Mutability of DNA base pairs: A statistical approach based on linear discrimination

Deborah L. HALL, Karen KAFADAR and Alvin M. MALKINSON

Statistical methodology for assessing homology of intronic regions of genes

Raymond J. CARROLL, Laurence S. FREEDMAN, Victor KIPNIS and Li LI

A new class of measurement error models, with applications to dietary data

S. DURHAM, N. FLUORNOY and W. LI

A sequential design for maximizing the probability of a favorable response

Frank HAMPEL

Is statistics too difficult?

Volume 26, No. 4, December/décembre 1998

Kiros BERHANE and Robert J. TIBSHIRANI

Generalized additive models for longitudinal data

J.V. ZIDEK, N.D. LE, H. WONG and R.T. BURNETT

Including structural measurement errors in the nonlinear regression analysis of clustered data

J.F. LAWLESS and M. ZHAN

Analysis of interval-grouped recurrent event data using piece-wise constant rate functions

Ming Gao GU and Shaolin LI

A stochastic approximation algorithm for maximum likelihood estimation with incomplete data

Jiahua CHEN

Penalized likelihood ratio test for finite mixture models with multinomial observations

E. SUSKO, J.D. KALBFLEISCH and J. CHEN

Constrained nonparametric maximum likelihood estimation for mixture models

Qiqing YU, Anton SCHICK, Linxiong LI and George Y.C. WONG

Asymptotic properties of the GMLE in the case 1 interval-censorship model with discrete inspection times

Claude BÉLISLE

Slow convergence of the Gibbs sampler

Kenny CRUMP and Daniel KREWSKI

Estimation of the number of studies with positive trends when studies with negative trends are present

Marlos A.G. VIANA, André ROGATKO and Timothy R. REBBECK

Statistical assessment of multi-valued screening tests

JOURNAL OF OFFICIAL STATISTICS

An International Review Published by Statistics Sweden

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 14, Number 2, 1998

The Sixth Morris Hansen Lecture: Opening Remarks <i>Daniel Levine</i>	115
The Hansen Era: Statistical Research and Its Implementation at the Census Bureau, 1940 - 1970 <i>Joseph Waksberg</i>	119
Discussion <i>Margo Anderson</i>	137
Discussion: The Irresistible Appeal of the Hansen Years at the U.S. Census Bureau <i>Robert M. Groves</i>	143
Sequential Poisson Sampling <i>Esbjörn Ohlsson</i>	149
Estimating the Sampling Variance of the UK Index of Production <i>P.N. Kokic</i>	163
Bias Correction in the Balanced-Half-Sampled Method if the Number of Sampled Units in Some Strata is Odd <i>Ger T. Slootbeek</i>	181
Response Burden and Panel Attrition <i>Adriaan W. Hoogendoorn and Dirk Sikkel</i>	189
London Plague Statistics in 1665 <i>David R. Bellhouse</i>	207
In Other Journals	235

Volume 14, Number 3, 1998

The Time Series Analysis of Compositional Data <i>Teresa M. Brunsdon and T.M.F. Smith</i>	237
Linking of Classifications by Linear Mappings <i>Beat Hulliger</i>	255
Determining Sample Sizes for Surveys with Data Analyzed by Hierarchical Linear Models <i>Michael P. Cohen</i>	267
Design and Analysis of Experiments Embedded in Sample Surveys <i>Jan van den Brakel and Robbert H. Renssen</i>	277
Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs <i>Lawrence R. Ernst</i>	297
Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling <i>Yves G. Berger</i>	315
Book and Software Reviews	325
In Other Journals	335

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

GUIDELINES FOR MANUSCRIPTS

Before having a manuscript typed for submission, please examine a recent issue (Vol. 19, No. 1 and onward) of *Survey Methodology* as a guide and note particularly the following points:

1. Layout

- 1.1 Manuscripts should be typed on white bond paper of standard size ($8\frac{1}{2} \times 11$ inch), one side only, entirely double spaced with margins of at least $1\frac{1}{2}$ inches on all sides.
- 1.2 The manuscripts should be divided into numbered sections with suitable verbal titles.
- 1.3 The name and address of each author should be given as a footnote on the first page of the manuscript.
- 1.4 Acknowledgements should appear at the end of the text.
- 1.5 Any appendix should be placed after the acknowledgements but before the list of references.

2. Abstract

The manuscript should begin with an abstract consisting of one paragraph followed by three to six key words. Avoid mathematical expressions in the abstract.

3. Style

- 3.1 Avoid footnotes, abbreviations, and acronyms.
- 3.2 Mathematical symbols will be italicized unless specified otherwise except for functional symbols such as “exp(·)” and “log(·)”, etc.
- 3.3 Short formulae should be left in the text but everything in the text should fit in single spacing. Long and important equations should be separated from the text and numbered consecutively with arabic numerals on the right if they are to be referred to later.
- 3.4 Write fractions in the text using a solidus.
- 3.5 Distinguish between ambiguous characters, (e.g., w, ω; o, O, 0; l, 1).
- 3.6 Italics are used for emphasis. Indicate italics by underlining on the manuscript.

4. Figures and Tables

- 4.1 All figures and tables should be numbered consecutively with arabic numerals, with titles which are as nearly self explanatory as possible, at the bottom for figures and at the top for tables.
- 4.2 They should be put on separate pages with an indication of their appropriate placement in the text. (Normally they should appear near where they are first referred to).

5. References

- 5.1 References in the text should be cited with authors' names and the date of publication. If part of a reference is cited, indicate after the reference, e.g., Cochran (1977, p. 164).
- 5.2 The list of references at the end of the manuscript should be arranged alphabetically and for the same author chronologically. Distinguish publications of the same author in the same year by attaching a, b, c to the year of publication. Journal titles should not be abbreviated. Follow the same format used in recent issues.

DIRECTIVES CONCERNANT LA PRÉSENTATION DES TEXTES

Avant de dactylographier votre texte pour le soumettre, prière d'examiner un numéro récent de *Techniques d'enquête* (à partir du vol. 19, n° 1) et de noter les points suivants:

1. Présentation

- 1.1 Les textes doivent être dactylographiés sur un papier blanc de format standard (8½ par 11 pouces), sur une face seulement, à double interligne partout et avec des marges d'au moins 1½ pouce tout autour.
- 1.2 Les textes doivent être divisés en sections numérotées portant des titres appropriés.
- 1.3 Le nom et l'adresse de chaque auteur doivent figurer dans une note au bas de la première page du texte.
- 1.4 Les remerciements doivent paraître à la fin du texte.
- 1.5 Toute annexe doit suivre les remerciements mais précéder la bibliographie.

2. Résumé

Le texte doit commencer par un résumé composé d'un paragraphe suivi de trois à six mots clés. Éviter les expressions mathématiques dans le résumé.

3. Rédaction

- 3.1 Éviter les notes au bas des pages, les abréviations et les sigles.
- 3.2 Les symboles mathématiques seront imprimés en italique à moins d'une indication contraire, sauf pour les symboles fonctionnels comme $\exp(\cdot)$ et $\log(\cdot)$ etc.
- 3.3 Les formules courtes doivent figurer dans le texte principal, mais tous les caractères dans le texte doivent correspondre à un espace simple. Les équations longues et importantes doivent être séparées du texte principal et numérotées en ordre consécutif par un chiffre arabe à la droite si l'auteur y fait référence plus loin.
- 3.4 Écrire les fractions dans le texte à l'aide d'une barre oblique.
- 3.5 Distinguer clairement les caractères ambigus (comme w, ω; o, O; 0; 1, I).
- 3.6 Les caractères italiques sont utilisés pour faire ressortir des mots. Indiquer ce qui doit être imprimé en italique en le soulignant dans le texte.

4. Figures et tableaux

- 4.1 Les figures et les tableaux doivent tous être numérotés en ordre consécutif avec des chiffres arabes et porter un titre aussi explicatif que possible (au bas des figures et en haut des tableaux).
- 4.2 Ils doivent paraître sur des pages séparées et porter une indication de l'endroit où ils doivent figurer dans le texte. (Normalement, ils doivent être insérés près du passage qui y fait référence pour la première fois).

5. Bibliographie

- 5.1 Les références à d'autres travaux faites dans le texte doivent préciser le nom des auteurs et la date de publication. Si une partie d'un document est citée, indiquer laquelle après la référence.
- 5.2 La bibliographie à la fin d'un texte doit être en ordre alphabétique et les titres d'un même auteur doivent être en ordre chronologique. Distinguer les publications d'un même auteur et d'une même année en ajoutant les lettres a, b, c, etc. à l'année de publication. Les titres de revues doivent être écrits au long. Suivre le modèle utilisé dans les numéros récents.

JOS is a scholarly quarterly that specializes in statistical methodology and applications. Survey methodology and other issues pertinent to the production of statistics at national offices and other statistical organizations are emphasized. All manuscripts are rigorously reviewed by independent referees and members of the Editorial Board.

Contents Volume 14, Number 2, 1998

The Sixth Morris Hansen Lecture: Opening Remarks	115
<i>Daniel Levine</i>	
The Hansen Era: Statistical Research and Its Implementation at the Census Bureau, 1940 - 1970	119
<i>Joseph Waksberg</i>	
Discussion	
<i>Margo Anderson</i>	137
Discussion: The Irresistible Appeal of the Hansen Years at the U.S. Census Bureau	
<i>Robert M. Groves</i>	143
Sequential Poisson Sampling	
<i>Esbjörn Ohlsson</i>	149
Estimating the Sampling Variance of the UK Index of Production	
<i>P.N. Kokic</i>	163
Bias Correction in the Balanced-Half-Sampled Method if the Number of Sampled Units in Some Strata is Odd	
<i>Ger T. Slootbeek</i>	181
Response Burden and Panel Attrition	
<i>Adrian W. Hoogendoorn and Dirk Sikkel</i>	189
London Plague Statistics in 1665	
<i>David R. Bellhouse</i>	207
In Other Journals	235

Volume 14, Number 3, 1998

The Time Series Analysis of Compositional Data	237
<i>Teresa M. Brunson and T.M.F. Smith</i>	
Linking of Classifications by Linear Mappings	
<i>Beat Hultigier</i>	255
Determining Sample Sizes for Surveys with Data Analyzed by Hierarchical Linear Models	
<i>Michael P. Cohen</i>	267
Design and Analysis of Experiments Embedded in Sample Surveys	
<i>Jan van den Brakel and Robert H. Kenssen</i>	277
Maximizing and Minimizing Overlap When Selecting a Large Number of Units per Stratum Simultaneously for Two Designs	
<i>Lawrence R. Ernst</i>	297
Variance Estimation Using List Sequential Scheme for Unequal Probability Sampling	
<i>Yves G. Berger</i>	315
Book and Software Reviews	325
In Other Journals	335

All inquiries about submissions and subscriptions should be directed to the Chief Editor:
Lars Lyberg, R&D Department, Statistics Sweden, Box 24 300, S - 104 51 Stockholm, Sweden.

Volume 26, No. 3, September/septembre 1998

Geoffrey WATSON

On the role of statistics in the paleomagnetic proof of continental drift

L. BILLARD

Some statistical opportunities in the agricultural sciences

Noel CRESSIE

Transect-spacing design of ice cores on the Antarctic continent

Daniel GERVINI and Victor J. YOHAI

Robust estimation of variance components

David R. BRILLINGER and Brent S. STEWART

Elephant seal movements: modelling migration

Phaik Mooi LEONG and Stephan MORGENTHAUER

Mutability of DNA base pairs: A statistical approach based on linear discrimination

Deborah L. HALL, Karen KAFADAR and Alvin M. MALKINSON

Statistical methodology for assessing homology of intronic regions of genes

Raymond J. CARROLL, Laurence S. FREEDMAN, Victor KIPNIS and Li LI

A new class of measurement error models, with applications to dietary data

S. DURHAM, N. FLUORNOY and W. LI

A sequential design for maximizing the probability of a favorable response

Frank HAMPEL

Is statistics too difficult?

Volume 26, No. 4, December/décembre 1998

Kiros BERHANE and Robert J. TIBSHIRANI

Generalized additive models for longitudinal data

J. V. ZIDEK, N. D. LE, H. WONG and R. T. BURNETT

Including structural measurement errors in the nonlinear regression analysis of clustered data

J. F. LAWLESS and M. ZHAN

Analysis of interval-grouped recurrent event data using piecewise constant rate functions

Ming Gao GU and Shaolin LI

A stochastic approximation algorithm for maximum likelihood estimation with incomplete data

Jiahua CHEN

Penalized likelihood ratio test for finite mixture models with multinomial observations

E. SUSKO, J. D. KALBFLEISCH and J. CHEN

Constrained nonparametric maximum likelihood estimation for mixture models

Qiqing YU, Anton SCHICK, Linxiong LI and George Y. C. WONG

Asymptotic properties of the GML in the case of interval-censorship model with discrete inspection times

Claude BÉLISLE

Slow convergence of the Gibbs sampler

Kenny CRUMP and Daniel KREWSKI

Estimation of the number of studies with positive trends when studies with negative trends are present

Marios A. G. VIANA, André ROGATKO and Timothy R. REBBECK

Statistical assessment of multi-valued screening tests

REMERCIEMENTS

Techniques d'enquête désire remercier les personnes suivantes, qui ont accepté de faire la critique d'un article durant l'année 1998. Un astérisque indique que la personne a participé plus d'une fois.

- * P. Lavallée, *Statistique Canada*
 J. Lawless, *University of Waterloo*
 G. Lee, *Australian Bureau of Statistics*
 S. Linacre, *Australian Bureau of Statistics*
 R. Little, *University of Michigan*
 S. Lohr, *Arizona State University*
 H. Mantel, *Statistique Canada*
 P. L. do Nascimento Silva, *IBGE, Brasil*
 G. Nathan, *Central Bureau of Statistics, Israel*
 D. Paton, *Statistique Canada*
 D. Pfeiffermann, *Hebrew University*
 N.G.N. Prasad, *University of Alberta*
 B. Quenneville, *Statistique Canada*
 T.E. Raghunathan, *University of Michigan*
 E. Rancourt, *Statistique Canada*
 J.N.K. Rao, *Carleton University*
 J.O. Ramsey, *McGill University*
 L.-P. Rivest, *Université Laval*
 G. Roberts, *Statistique Canada*
 G. Robinson, *Commonwealth Scientific and Industrial Research Organisation*
 I. Sande, *Bell Communications Research, U.S.A.*
 C.-E. Sarnadal, *Université de Montréal*
 N. Schenker, *University of California - Los Angeles*
 F.J. Scheuren, *George Washington University*
 A.J. Scott, *University of Auckland*
 J. Sedransk, *Case Western Reserve University*
 M.P. Singh, *Statistique Canada*
 R. Sitter, *Simon Fraser University*
 C.J. Skinner, *University of Southampton*
 K.P. Srinath, *ABT Associates, Inc.*
 L. Stokes, *University of Texas - Austin*
 D. Stukel, *Statistique Canada*
 A. Théberge, *Statistique Canada*
 S. Thivierge, *Statistique Canada*
 R. Thomas, *Carleton University*
 I. Thomsen, *Statistics Norway*
 R. Tortora, *Gallup Organization*
 C. Tucker, *U.S. Bureau of Labor Statistics*
 R. Valliant, *U.S. Bureau of Labor Statistics*
 V.K. Verma, *University of Essex*
 P.J. Waite, *U.S. Bureau of the Census*
 J. Waksberg, *Westat, Inc.*
 K.M. Wolter, *National Opinion Research Center*
 F. Yu, *Australian Bureau of Statistics*
 E. Zanutto, *University of Pennsylvania*
 A. Zaslavsky, *Harvard University*
- * B.M. Balk, *Statistics Netherlands*
 D.R. Bellhouse, *University of Western Ontario*
 J. Bethel, *Westat, Inc.*
 P. Biemer, *Research Triangle Institute*
 D.A. Binder, *Statistique Canada*
 J.-R. Boudreau, *Statistique Canada*
 K. Brewer, *Australian National University*
 J.M. Brick, *Westat, Inc.*
 A. Brinkley, *U.S. Bureau of the Census*
 T.W. Broene, *Energy Information Administration, U.S.A.*
 N. Buck, *University of Essex*
 P. Cantwell, *U.S. Bureau of the Census*
 R. Chambers, *University of Southampton*
 M. Cohen, *National Center for Education Statistics*
 B. Cox, *Mathematica Policy Research*
 J. Denis, *Statistique Canada*
 J.-C. Deville, *Institut national de la statistique et des études économiques*
 P. Dick, *Statistique Canada*
 A. Dorfman, *U.S. Bureau of Labor Statistics*
 J.D. Drew, *Statistique Canada*
 J. Dufour, *Statistique Canada*
 * J. Eltinge, *Texas A&M University*
 R. Evans, *University of Nebraska - Lincoln*
 M. Feder, *University of Southampton*
 W.A. Fuller, *Iowa State University*
 * J. Gambino, *Statistique Canada*
 B. Graubard, *U.S. National Cancer Institute*
 * R.M. Groves, *University of Maryland*
 S.J. Haslett, *Massey University, New Zealand*
 L. Hattersley, *Office for National Statistics, U.K.*
 * M.A. Hidiroglou, *Statistique Canada*
 * D. Holt, *Office for National Statistics, U.K.*
 K. Humphreys, *University of Glasgow*
 J.-S. Hwang, *Academia Sinica*
 W. Jocelyn, *Statistique Canada*
 E. Johnson, *Educational Testing Service, U.S.A.*
 D. Judkins, *Westat, Inc.*
 * G. Kalton, *Westat, Inc.*
 B. King, *Case Western Reserve University*
 * P.S. Kott, *National Agricultural Statistics Service*
 M. Kovachevic, *Statistique Canada*
 M. Kramer, *U.S. Bureau of the Census*
 A. Krieger, *University of Pennsylvania*
 * R. Lachapelle, *Statistique Canada*
 * P. Lahiri, *University of Nebraska - Lincoln*
 M. Larson, *Harvard University*

On remercie également ceux qui ont contribué à la production des numéros de la revue pour 1998: J. Beauseigle (Division de la diffusion) et L. Perreault (Division des langues officielles et traduction). Finalement on désire exprimer notre reconnaissance à D. Blair, C. Larbie, D. Lemire, C. Marleau et G. Ray de la Division des méthodes des enquêtes auprès des ménages, pour leur apport à la coordination, la dactylographie et la rédaction.

KORN, E.L., et GRAUBARD, B.T. (1998). Variance estimation for superpopulation parameters. *Statistica Sinica*, 8, 1131-1151.

KOVAR, M.G., et JOHNSON, C. (1986). Design effects from the Mexican American portion of the Hispanic Health and Nutrition Examination Survey: a strategy for analysts. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-399.

KREWSKI, D., et RAO, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods. *Annals of Statistics*, 9, 1010-1019.

McQUILLAN, G.M., KHARE, M., KARON, J.M., SCHABLE, C.A., et VLAHOV, D. (1997). Update on the seroepidemiology of Human Immunodeficiency Virus in the United States household population: NHANES III, 1988-94. *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, 14, 355-360.

NATIONAL CENTER FOR HEALTH STATISTICS (1985). Plan and operation of the Hispanic Health and Nutrition Examination Survey, 1982-84. *Vital and Health Statistics* 1(19). Hyattsville, MD: National Center for Health Statistics.

VOLLSET, S.E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, 12, 809-824

NATIONAL CENTER FOR HEALTH STATISTICS (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988-94. *Vital and Health Statistics*, 1(32). Hyattsville, MD: National Center for Health Statistics.

PFEFFERMAN, D., et LAVANGE, L. (1989). Regression models for stratified multi-stage cluster samples. Dans *Analysis of Complex Surveys*, (Eds. C.J. Skinner, D. Holt, et T.M.F. Smith). New York: Wiley, 237-260.

SAMHSA (1998). National Household Survey on Drug Abuse: Main Findings 1996. (DHHS Publication No. (SMA) 98-3200). Rockville, MD: SAMHSA.

SAS (1990). *SAS Language: Reference, Version 6*. Première édition. Cary, NC: SAS Institute Inc.

SHAH, B.V., BARNWELL, B.G., et BIELER, G.S., (1995). *SUDAAN User's Manual, Release 6.40*. Research Triangle Park, NC: Research Triangle Institute.

croyons que l'effet véritable du plan de sondage dépasse l'effet véritable de la pondération de l'échantillon. (Il est difficile d'estimer l'effet du plan de sondage étant donné le nombre restreint de degrés de liberté.) C'est pourquoi les méthodes de troncation paraissent raisonnables pour cette application, à notre avis. En raison des degrés de liberté peu nombreux et parce que l'usage de la cocaïne n'est pas rare, on note plus de variations entre les intervalles de confiance à logit, de Breeze et proposés, au tableau 7. Les simulations décrites à la partie 3 nous portent à recommander les intervalles de confiance proposés (avec troncation). Notre approche pourrait donner l'impression d'être légèrement incohérente dans le cas de cette enquête, en ce sens que nous tolérons des échantillons d'une taille efficace estimée inférieure à la taille de l'échantillon, et les tronquons davantage. Nous pensons toutefois qu'il s'agit d'une approche raisonnablement conservatrice lorsqu'on soupçonne que l'effet véritable du plan de sondage dépasse l.

5. DISCUSSION

Quoique les intervalles de confiance proposés présentent une couverture probable adéquate pour toutes les simulations ou presque, on ne peut garantir ce résultat pour l'ensemble des configurations de la population (voir le dernier tiers du tableau 5). Il est facile d'illustrer une non-couverture plus importante. Supposons, par exemple, que la population se compose de grappes de 100 éléments et que 10 % des grappes ne renferment que des éléments positifs et les 90 % restants, aucun. En prélevant 10 grappes par échantillonnage aléatoire simple et en prenant toutes les unités des grappes retenues, dans 35 % (= $(1-0,1)^{10}$) des cas, l'échantillon de 1 000 ne comprendra aucune unité positive. Dans une situation de ce genre, les intervalles proposés se résument aux intervalles binomiaux usuels, c'est à-dire la limite supérieure de l'intervalle de confiance de 95 % est $0,003 (= 1-0,05^{1/1000})$. Il s'ensuit que la limite supérieure de l'intervalle de confiance de 95 % est inférieure à la valeur véritable de 0,10 dans au moins 35 % des cas, signe d'une sérieuse non-couverture.

Dans les cas d'échantillonnage simples, on peut bâtir des intervalles de confiance qui respectent au moins la probabilité de couverture nominale en tenant compte de tous les agencements possibles de population et en retenant la configuration la moins favorable pour calculer la couverture probable. Dans le cas hypothétique de l'échantillon à un degré mentionné ci-dessus, par exemple, la limite binomiale calculée à partir de 0 unité positive sur 10, soit 0,26 (= $1-0,05^{1/10}$) pourrait servir de limite supérieure à l'intervalle de confiance de 95 %. Gross et Frankel (1991) ont étudié de tels intervalles, dont le calcul peut devenir très laborieux, et suggèrent eux aussi des approximations réclamant des opérations moins nombreuses.

Les auteurs désirent remercier M. Khare pour leur avoir fourni les estimations sur la prévalence et l'erreur-type attribuable au plan de sondage qui s'y rapporte apparaissant au tableau 6, ainsi que le rédacteur associé et les examinateurs pour leurs précieux commentaires.

REMERCIEMENTS

Comparativement aux approches de ce genre, les intervalles proposés ont pour avantage (1) d'être faciles à calculer, (2) de se plier aux plans d'échantillonnage complexes, y compris à l'ajustement des poids pour la non-réponse et par stratification a posteriori, (3) de généralement garder leur probabilité de couverture nominale, (4) d'être moins conservateurs que les intervalles qui préservent leur couverture nominale sans égard à la configuration de la population et (5) de présenter des propriétés plus intéressantes que les intervalles linéaires, à logit ou de Breeze. Les conclusions (2) et (5) reposent sur les résultats des simulations qui, il va de soi, ne couvrent pas toutes les situations imaginables. Il conviendrait d'entreprendre d'autres recherches en la matière.

AGRESTI, A. (1990). *Categorical Data Analysis*. New York: Wiley.

BREEZE, E. (1990). *General Household Survey: Report on Sampling Error*. London: Her Majesty's Stationery Office (Office of Population Censuses and Surveys).

CLOPPER, C.J., et PEARSON, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

COCHRAN, W.G. (1977). *Sampling Techniques*. Troisième édition. New York: Wiley.

FRANKEL, M.R. (1971). *Inference from Survey Samples: An Empirical Investigation*. Ann Arbor, MI: Institute for Social Research, The University of Michigan.

GRAUBARD, B.I., et KORN, E.L. (1996). Modelling the sampling design in the analysis of health surveys. *Statistical Methods in Medical Research*, 5, 263-281.

GROSS, S.T., et FRANKEL, M.R. (1991). Confidence limits for small proportions in complex samples. *Communications in Statistics - Theory and Methods*, 20, 951-975.

GÜENTHER, W.C. (1983). Hypergeometric distributions. Dans *Encyclopedia of Statistical Sciences*. Volume 3, (Eds. S. Kotz et N.L. Johnson). New York: Wiley, 707-712.

JENNINGS, D.E. (1987). How do we judge confidence-interval adequacy? *American Statistician*, 41, 335-337.

JOHNSON, N.L., et KOTZ, S. (1970). *Continuous Univariate Distributions - 2*. New York: Wiley.

JOHNSON, N.L., KOTZ, S., et KEMP, A.W. (1993). *Univariate Discrete Distributions*. Deuxième édition. New York: Wiley.

KORN, E.L., et GRAUBARD, B.I. (1990). Simultaneous testing or regression coefficients with complex survey data: use of Bonferroni t-statistics. *American Statistician*, 44, 270-276.

et la nutrition

	Taille de l'échantillon	Nombre infectés	Prévalence (%) ± ET	Taille efficace de l'éch.	n*	n _{df} *
Total	1 1202	59	0,320 ± 0,076	0,519 ± 0,130	0,127 ± 0,053	0,203 ± 0,071
Masculin	5142	44				
Fémnin	6060	15				
Race/ascesdance						
Blanc	4128	9				
Noir	3579	38				
Mex.-Amér.	3495	12				

Tableau 7

Sexe	Total	Masculin	Féminin
Taille de l'échantillon	123	69	54
A essayé	13	10	3
Proportion (%) \pm ET	11,6 \pm 2,5	14,3 \pm 3,4	7,0 \pm 4,8
n^*	1671	1050	282
n^{df}	1328	844	229
Int. conf. linéaire 90 %	(7,0, 16,2)	(8,0, 20,7)	(-1,9 ^a , 15,9)
Int. conf. logit 90 %	(7,8, 17,1)	(9,1, 21,9)	(1,9, 22,8)
Int. conf. Brezée 90 %	(7,7, 17,0)	(8,3, 23,2)	(0,9, 24,8)
Int. conf. proposé 90 %	(7,4, 17,2)	(8,5, 22,1)	(0,9, 22,7)
Avec troncation			
Int. conf. linéaire 90 %	(6,3, 17,0)	(6,5, 22,2)	voir ci-dessus
Int. conf. logit 90 %	(7,2, 18,2)	(8,1, 24,1)	“
Int. conf. Brezée 90 %	(7,1, 18,1)	(7,7, 24,4)	“
Int. conf. proposé 90 %	(7,2, 17,5)	(8,0, 23,2)	“

4.2 Cocaïnomanie chez les diplômés de collège
échantillonnés lors de l'Enquête sur la santé et la
nutrition des personnes d'origine hispanique
(HHANES)

La HHANES s'est déroulée en 1982-1983 et portait sur trois groupes hispanophones des Etats-Unis (National Center for Health Statistics 1985). Nous ne nous sommes intéressés qu'à l'échantillon Mexicain-Américain. On a demandé aux personnes de 12 à 44 ans à quel âge elles avaient essayé de la cocaïne pour la première fois. Les réponses possibles étaient l'âge du répondant lorsqu'il avait essayé de la cocaïne pour la première fois (en années), «n'a jamais essayé» et «ne sait pas». Nous avons tenté d'estimer la proportion de ceux qui avait déjà fait l'expérience de la

drogue parmi les personnes comptant 16 années ou plus de scolarité (pour lesquelles la réponse «ne sait pas» n'existait pas). Sur les 123 personnes échantillonnées, 13 avait déjà absorbé de la cocaïne, soit une proportion de 11,6 % après pondération (tableau 7). L'erreur-type attribuable au plan de sondage (2,5 %) n'est estimée qu'avec 8 degrés de liberté parce qu'on peut reproduire approximativement le plan de sondage de l'enquête en prélevant 2 UPE dans chacune des 8 strates (Kovar et Johnson 1986). La taille efficace de l'échantillon s'établit à $n^* = 167,1$ et $n_d^* = 132,8$, soit plus que l'échantillon, dans les deux cas. On le doit à l'effet estimé du plan de sondage (0,73, de sorte que $n^* = 123/0,736 = 167,1$). (Le deuxième facteur de (2.2) est égal à 0,794.) Malgré la stratification, nous

rend compte que les limites proposées créent un intervalle de 4,0 % plus large avec troncation, comparativement à sans (résultats non indiqués pour les limites tronquées).

Le tableau 4 montre les résultats d'une simulation identique à celle du tableau 1, sauf pour le nombre d'observations par grappe, qui passe de 100 à 10. Les résultats ressemblent fort à ceux du tableau 1 pour les simulations comprenant le même nombre prévu d'événements positifs. Seuls les intervalles de Breeze, plus conservateurs que ceux de la méthode préconisée, font exception à la règle. On le doit au fait que la proportion globale est plus importante au tableau 4 qu'au tableau 1, pour le même nombre d'événements prévus (l'échantillon du tableau 4 étant plus petit). Les intervalles de Breeze modifiés par l'approximation de Poisson ne donnent pas de bons résultats avec les grandes proportions. Pour le vérifier, nous avons effectué une simulation identique à celle du tiers supérieur du tableau 1 mais en prenant $p = 0,5$ comme proportion pour 1 600 événements positifs prévus. Les pourcentages inférieur et supérieur de non-couverture étaient de 1,2 % et 1,3 %, respectivement, contre 4,6 % et 4,7 % avec la méthode proposée. En moyenne, les intervalles de Breeze sont 37 % plus larges que les intervalles proposés.

Les intervalles de Breeze laissent aussi à désirer avec un très petit nombre de grappes, car ils négligent le nombre de degrés de liberté dans l'estimation de la variance. Nous avons, par exemple, effectué une simulation correspondant au premier tiers du tableau 1 en ne prenant les données que de 8 grappes (100 observations par grappe), et $p_i = 0,1$, de telle sorte que le nombre d'événements positifs prévus se situait à 80. Les pourcentages inférieur et supérieur de non-couverture s'établissaient à 6,1 % et à 5,4 %, respectivement, pour les intervalles de Breeze, comparativement à 4,7 % et à 4,0 % pour la méthode préconisée.

Le tableau 5 donne les résultats d'une simulation identique à celle du tableau 1, mais avec des grappes de 10 ou de 100, selon une probabilité de 1/2. Les probabilités de non-couverture dépassent la valeur nominale de 5 % du tiers inférieur du tableau, peu importe la méthode employée. En outre, les intervalles à logit ne se comportent pas aussi bien qu'au tableau 1 dans les deux premiers tiers du tableau. Les résultats (non indiqués) se rapprochaient beaucoup de ceux du tableau 1.

4. APPLICATIONS

Nous examinerons maintenant deux applications caractérisées par un petit nombre d'événements positifs. La

première concerne l'estimation de l'infection par le VIH au sein d'une population non sélectionnée. Le nombre de réagissants est peu élevé, car le taux d'infection par le VIH est relativement faible. Dans le deuxième cas (estimation de la cocaïnomanie), les taux sont élevés mais le nombre d'événements positifs reste faible parce que l'analyse ne s'applique qu'à de petits sous-groupes. Dans les deux cas, on a utilisé SUDAAN (Shah, Barnwell et Bieler 1995) pour calculer l'erreur-type (attribuable au plan de sondage) des proportions et la fonction «FINV» du SAS (1990, p. 547) pour établir les quantiles de la distribution F en (1.2).

4.1 Estimation de la séroprévalence du VIH dans le cadre de la troisième Enquête nationale sur la santé et la nutrition (NHANES III)

L'enquête NHANES III s'est déroulée de 1988 à 1994 et couvrait les membres de la population civile non institutionnalisés de 2 mois et plus aux États-Unis (National Center for Health Statistics 1994). Les participants de 18 ans et plus ont subi un test de dépistage du VIH. McQuillan, Khare, Karon, Schable et Vlahov (1997) se sont penchés sur la prévalence du virus chez les personnes de moins de 60 ans et dans divers sous-groupes, dont certains sont reproduits au tableau 6. Sur les 11 202 personnes testées, 59 portaient le virus. La prévalence estimée au tableau 6 (0,32 %) s'écarte considérablement de la proportion non pondérée (0,53 % = 59/11 202) parce qu'elle correspond à une proportion pondérée par les poids de l'échantillon. Les tests de dépistage préservant l'anonymat, les poids d'échantillonnage ont été dérivés des poids originaux de l'échantillon de la NHANES III pour les personnes testées présentant les mêmes caractéristiques, soit mêmes flots (lieu de l'enquête), race ou groupe ethnique, sexe et groupe d'âge (18-39 et 40-59 ans) (communication personnelle de M. Khare). Le pseudo-plan de sondage utilisé pour estimer la variance consistait à prélever 2 pseudo-UPB dans chacune des 23 strates (communication personnelle de M. Khare), ce qui ne correspond pas au pseudo-plan de sondage typique servant à estimer la variance dans la NHANES III.

Les intervalles de confiance linéaires de 90 % s'appliquant à la prévalence du VIH dans les différents groupes, au tableau 6, sont décalés à gauche et sont plus étroits que les autres intervalles, qui se ressemblent. Les intervalles proposés sont légèrement plus larges que ceux de Breeze ou à logit. La taille efficace de l'échantillon établie en (2.1) est nettement plus faible que celle de l'échantillon, à cause des effets du plan de sondage; les intervalles de confiance obtenus après troncation seront donc identiques à ceux du tableau 6. L'écart entre n^* et n_{df} s'avère relativement faible. Pour cet événement positif relativement rare, les simulations décrites à la partie 3 donnent à penser que les intervalles de confiance de Breeze et ceux proposés maintiennent mieux la probabilité de couverture nominale de 90 % que les autres intervalles.

Tableau 5
Simulation de la non-couverture (en pour cent) des limites inférieure et supérieure d'un intervalle de confiance unilatéral de 95 % pour un échantillon de 32 grappes de 10 ou de 100 observations chacune avec probabilité de 1/2; poids d'échantillonnage de 1 ou 10 avec probabilité de 1/2 (poids non informatifs)

Distribution des proportions pour globale	Proportion	Nombre	Méthode de calcul des limites de l'intervalle			
			Linéaire		Logit	
proportions pour globale			Sup.	Infér.	Sup.	Infér.
les grappes ^a						
proposée						

0,1818	0,1818	320	5,1	6,0	5,7	5,2	4,2	4,1	5,2	5,0
0,0364	0,0364	64	4,1	7,6	5,7	5,0	4,7	5,2	4,8	4,9
0,0182	0,0182	32	3,4	8,5	5,7	5,0	4,7	5,1	4,4	4,7
0,0045	0,0045	8	2,0	12,7	5,9	3,4	4,0	4,3	3,6	3,8
(1/2, 1/2)	0,0909, 0,2727	0,1818	5,0	6,4	6,1	4,8	4,2	3,6	5,2	4,4
	0,0182, 0,0545	0,0364	3,9	8,1	6,0	5,1	4,9	5,0	4,7	4,8
	0,0091, 0,0273	0,0182	3,1	9,3	5,8	5,2	4,5	5,3	4,2	4,9
	0,0023, 0,0068	0,0045	1,8	13,2	5,9	3,6	3,9	4,5	3,5	4,0
(3/4, 1/4)	0,0909, 0,4545	0,1818	3,1	9,9	4,6	7,6	2,5	6,3	3,3	7,1
	0,0182, 0,0909	0,0364	2,8	10,9	5,3	7,3	3,9	7,3	3,7	7,0
	0,0091, 0,0455	0,0182	2,4	11,5	5,4	6,8	3,9	6,9	3,6	6,5
	0,0023, 0,0114	0,0045	1,6	14,5	5,7	4,0	3,7	5,0	3,3	4,4
(a) Les fractions entre parenthèses indiquent la probabilité que la proportion de grappes pondérée corresponde à la valeur mentionnée.										

L'échantillon sont informatifs. Pour cela, on a fixé le poids à 10 avec probabilité de 2/3 si l'évènement était positif et de 1/3 dans le cas contraire, sinon le poids était de 1. La probabilité que chaque grappe comprenne un événement positif a été ajustée à la baisse jusqu'à ce que les proportions pondérées générales soient identiques à celles du tableau 1. Les résultats du tableau 2 ressemblent à ceux du tableau 1 mais les intervalles linéaire et à logit ont tendance à donner de moins bonnes probabilités de couverture.

Au tableau 3, on peut voir les résultats d'une simulation identique à celle du tableau 1, à l'exception du fait que l'analyse n'est pas pondérée. Comme le tiers supérieur du tableau 3 n'indique aucune corrélation entre les éléments des grappes, on pourrait aussi bien recourir aux limites binomiales de l'échantillon aléatoire simple. Si on effectue la moyenne des quatre possibilités dans ce tiers du tableau, on constate que les limites proposées donnent une largeur de 2,5 % supérieure à celle de l'intervalle binomial (résultats non indiqués). Puisque l'effet véritable du plan de sondage dans le tiers supérieur du tableau 3 est égal à 1,0, les simulations permettent d'examiner les conséquences de la troncation de n_{gf}^* dans la méthode proposée. (Il y a rarement troncation dans les simulations du tableau 1, car l'incidence véritable du plan de sondage est toujours >1.) Lorsqu'on applique la méthode suggérée avec troncation, la simulation donne des intervalles plus larges et plus conservateurs que ceux du premier tiers du tableau 3. Après calcul de la moyenne des quatre cas, on se

La limite supérieure de l'intervalle de confiance linéaire tombe plus en-dessous de sa valeur véritable que le niveau nominal de 5 %, même avec un compte prévu de jusqu'à 320 événements positifs, en particulier quand il y a une corrélation positive dans la grappe (tiers intermédiaire et inférieur du tableau), ce qui est quelque peu étonnant. Dans le cas de l'intervalle de confiance de la transformation par logit, la non-couverture semble dépasser légèrement le niveau nominal, surtout pour les limites inférieures. Les limites des intervalles de Breeze et des intervalles proposés paraissent généralement conservatrices. Celles des intervalles binomiaux avec échantillonnage aléatoire simple ne conviennent pas aux cas simulés au tableau 1, à cause des poids d'échantillonnage et de la corrélation dans la grappe (dans les deux tiers inférieurs du tableau). Une non-couverture de plus de 8 % pour les limites inférieure et supérieure de l'intervalle binomial pour tous les cas du tableau, en est la preuve (résultats non indiqués).

Discuter de la «longueur» des intervalles de confiance unilatéraux s'avère un peu compliqué aussi nous restreindrons-nous à parler de la longueur des intervalles bilatéraux à 90 %. Pour l'ensemble des simulations du tableau 1, les intervalles de Breeze et ceux que nous proposons ont une largeur moyenne de 3,3 % et de 4,9 % supérieure à celle des intervalles de transformation par logit.

Le tableau 2 présente les résultats d'une simulation semblable à celle du tableau 1, si ce n'est que les poids de

Tableau 3

Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance unilatéral de 95 % pour un échantillon de 32 grappes de 100 observations chacune; analyses non pondérées

Distribution des proportions pour les grappes ^a	Proportion globale prévue	Nombre	Méthode de calcul des limites de l'intervalle			
			Linéaire	Logit	Breze	Proposée
			Infér.	Sup.	Infér.	Sup.

(1)	0,1	320	5,0	4,9	5,7	4,2	4,9	3,8	5,2	4,1	4,4	4,0
0,02	0,02	64	3,8	6,3	5,2	4,5	4,7	4,8	4,4	4,3	4,4	4,4
0,01	0,01	32	3,5	6,8	5,6	4,4	4,7	4,4	4,3	4,3	4,0	3,9
0,0025	0,0025	8	2,5	8,8	5,6	3,8	4,1	3,9	3,9	3,9	3,9	3,9

(1/2, 1/2)	0,05, 0,15	320	4,5	5,6	5,6	4,2	4,5	3,7	4,8	4,0	4,6	4,4
0,05, 0,25	0,1	320	3,3	7,7	4,8	5,6	3,5	5,1	3,7	5,3	4,9	5,3
0,01, 0,05	0,02	64	2,9	8,1	5,1	5,2	4,1	5,3	3,8	4,9	4,9	5,2
0,005, 0,025	0,01	32	2,5	9,2	4,9	5,6	3,9	5,6	3,5	5,2	4,9	5,1
0,00125, 0,00625	0,0025	8	2,0	10,4	5,3	5,1	3,8	5,1	3,3	5,1	4,3	5,1

(a) Les fractions entre parenthèses indiquent la proportion de grappes correspondre à la valeur mentionnée.

Tableau 4

Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance unilatéral de 95 % pour un échantillon de 32 grappes de 10 observations chacune; poids d'échantillonnage de 1 ou 10 avec probabilité de 1/2 (poids non informatifs)

Distribution des proportions pour les grappes ^a	Proportion globale prévue	Nombre	Méthode de calcul des limites de l'intervalle			
			Linéaire	Logit	Breze	Proposée
			Infér.	Sup.	Infér.	Sup.

(1)	0,2	64	4,0	6,6	5,2	4,7	3,1	4,7	4,3	4,0	2,4	4,3
0,1	0,1	32	3,2	7,8	5,3	4,4	3,6	3,8	3,9	4,0	2,4	4,3
0,025	0,025	8	1,7	10,2	5,5	2,1	3,4	2,1	3,2	2,4	2,4	4,3

(1/2, 1/2)	0,1, 0,3	64	3,6	7,0	5,0	4,9	2,8	3,4	3,9	4,4	2,5	4,4
0,1, 0,15	0,1	32	3,0	8,1	5,1	4,6	3,4	4,0	3,7	4,2	2,5	4,4
0,0125, 0,0375	0,025	8	1,6	10,6	5,4	2,1	3,3	2,1	3,1	2,5	2,5	4,4

(3/4, 1/4)	0,1, 0,5	64	3,1	7,8	4,6	5,3	2,4	3,9	3,3	4,8	2,8	4,8
0,05, 0,25	0,1	32	2,5	9,2	4,8	5,2	3,0	4,6	3,3	4,8	2,8	4,8
0,0125, 0,0625	0,025	8	1,5	11,5	5,3	2,4	3,2	3,5	3,0	2,8	2,8	4,8

(a) Les fractions entre parenthèses indiquent la proportion de grappes correspondre à la valeur mentionnée.

Tableau 1

Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance unilatéral de 95 % comptant 32 grappes de 100 observations chacune; poids d'échantillonnage de 1 ou de 10 avec probabilité de 1/2 (poids non informatifs)

Distribution des proportions pour les grappes ^a	Proportion globale	Nombre prévu positif	Linéaire	Logit	Breeze	Proposée
			Infér.	Infér.	Infér.	Sup.
			Sup.	Sup.	Sup.	Sup.

(1)	0,1	320	4,6	5,5	4,6	4,8
	0,02	64	3,4	5,2	4,5	4,2
	0,01	32	2,9	5,4	4,5	4,0
	0,0025	8	1,6	5,5	3,6	3,3
0,0025						1,8

(1/2, 1/2)	0,05, 0,15	320	4,3	5,5	4,3	4,7
	0,01, 0,03	64	3,1	5,2	4,3	4,0
	0,005, 0,015	32	2,7	5,2	4,1	3,7
	0,00125, 0,00375	8	1,5	5,4	3,4	3,1

(3/4, 1/4)	0,05, 0,25	320	3,1	4,7	3,4	3,6
	0,01, 0,05	64	2,7	5,1	4,0	3,7
	0,005, 0,025	32	2,2	5,0	3,7	3,3
	0,00125, 0,00625	8	1,3	5,3	3,3	3,0

(a) Les fractions entre parenthèses indiquent la proportion de grappes correspondante à la valeur mentionnée.

Tableau 2

Simulation de la non-couverture (en pour cent) des limites supérieure et inférieure d'un intervalle de confiance de 95 % pour un échantillon de 32 grappes de 100 observations chacune; les poids d'échantillonnage sont informatifs et correspondent à 1 ou 10 (voir le corps du texte)

Distribution des proportions pour les grappes ^a	Proportion globale pondérée	Nombre prévu positif	Linéaire	Logit	Breeze	Proposée
			Infér.	Infér.	Infér.	Sup.
			Sup.	Sup.	Sup.	Sup.

(1)	0,1	191,0	4,3	5,9	4,9	4,6
	0,02	36,9	3,3	7,3	4,3	4,1
	0,01	18,4	2,8	8,7	4,0	3,9
	0,0025	4,6	1,3	18,7	3,2	2,8

(1/2, 1/2)	0,05, 0,15	191,0	5,0	6,4	3,7	5,4
	0,01, 0,03	36,9	3,0	7,9	4,5	4,0
	0,005, 0,015	18,4	2,5	9,2	4,2	3,7
	0,00125, 0,00375	4,6	1,3	19,0	3,2	2,8

(3/4, 1/4)	0,05, 0,25	191,0	4,7	5,7	4,1	5,5
	0,01, 0,05	36,9	2,6	8,9	5,2	3,7
	0,005, 0,025	18,4	2,1	10,1	4,8	3,4
	0,00125, 0,00625	4,6	1,2	19,8	3,3	2,8

(a) Les fractions entre parenthèses indiquent la proportion de grappes pondérée correspondante à la valeur mentionnée.

confiance de la transformation par logit (Agresti 1990, pp. 249-250). Dans le cas présent, quand $\hat{p} = 0$, l'intervalle de confiance a été fixé à l'intervalle binomial $(p^L(0, n), p^U(0, n))$.

Lorsqu'on sait que l'effet (véritable) du plan d'échantillonnage dépassera 1 avant l'échantillonnage, il est possible d'apporter diverses modifications aux procédures qui précèdent. Nous préconisons pour notre part de tronquer à n la taille efficace de l'échantillon ajustée en fonction du nombre de degrés de liberté. Bref, si n_{df}^* est

plus grand que n , nous en fixons la valeur à n et prenons comme limites supérieure et inférieure de l'intervalle $p^L(\hat{p}n, n)$ et $p^U(\hat{p}n, n)$. Pour les intervalles de Brezee, on pourrait établir que n^* est égal à n si $n^* > n$. Avec l'intervalle linéaire ou à logit, il est possible de prendre l'estimateur de la variance de l'échantillon aléatoire simple $\hat{p}(1 - \hat{p})/n$ au lieu de $\text{var}(\hat{p})$ dans (1.1), (2.3) et (2.4) si

$n^* > n$; on trouvera d'autres tronctions dans SAMHSA (1998). On justifie ces tronctions par le fait que l'instabilité de l'estimateur de la variance $\text{var}(\hat{p})$ peut donner un effet de plan estimé inférieur à un. L'instabilité en question peut s'avérer particulièrement importante puisque \hat{p} est petit (SAMHSA 1998). Ces méthodes de troncation ont pour conséquence d'élargir les intervalles de confiance et de les rendre plus conservateurs. En théorie, on pourrait aussi ajuster la taille réelle estimée de l'échantillon si on sait que l'effet (véritable) du plan de sondage est inférieur à un avant l'échantillonnage. Par mesure de prudence, nous recommandons toutefois de l'éviter.

Dans cet article, nous nous concentrons sur les intervalles de confiance de la probabilité de «superpopulation» que $Y = 1$ et n'égal pas la proportion de la population finie. En d'autres termes, le paramètre visé est $P = \sum_{N=1}^N p^N/N$ et non pas $P = \sum_{N=1}^N X^N/N$, où X^N présente une distribution de Bernoulli de paramètre p^N , et N correspond à la taille de la population. Les probabilités de couverture obtenues par simulation qui apparaissent à la section suivante se rapportent donc à la couverture de p . Ce paramètre étant établi, nous ne recourons pas aux facteurs de correction de la population finie au moment d'estimer $\text{var}(\hat{p})$ utilisé dans (2.2); nous ne procéderons pas ici à d'autres ajustements de la variance $\text{var}(\hat{p})$ attribuable au plan d'échantillonnage pour l'inférence à la superpopulation (Korn et Graubard 1998). Un examinateur a suggéré qu'on recoure à un modèle pour estimer l'intervalle de confiance de p . Bien que limites, nos essais indiquent néanmoins qu'une approche de ce genre donne des estimateurs semblables aux estimateurs pondérés et ne présente aucun avantage sur le plan de l'inférence (Pfeffermann et Lavange 1989; Graubard et Korn 1996).

Ceux qui s'intéressent à un intervalle de confiance pour P devraient se servir des intervalles proposés en intégrant des facteurs de correction pour la population finie à $\text{var}(\hat{p})$, dans (2.2). On pourrait établir un intervalle de confiance

pour $\sum_{N=1}^N X^N$ en multipliant les extrêmes de l'intervalle de P par N , s'il est connu, ou par l'estimateur \hat{N} de N , dans le cas contraire. (Théoriquement, on pourrait aussi tenir compte de la variabilité de \hat{N} , mais la variabilité supplémentaire sera faible.) Une autre façon d'estimer l'intervalle de confiance de P consisterait à modifier les limites usuelles (Guenther 1983) d'un échantillon aléatoire simple (d'après la distribution hypergéométrique), un peu comme les intervalles suggérés modifient les intervalles binomiaux.

3. SIMULATIONS

Les principaux résultats de la simulation apparaissent aux tableaux 1 à 5. Le tableau 1 donne les résultats des simulations pour 32 grappes de 100 unités chacune. Dans la grappe i , on a simulé le nombre d'événements positifs au moyen d'une distribution binomiale assortie de p_i comme paramètre probabiliste. Les valeurs $\{p_i, i = 1, \dots, 32\}$ du tableau 1 correspondent à la probabilité dans les grappes. Dans le tiers supérieur du tableau, cette probabilité est égale à la constante $p = 0,1, 0,02, 0,01$ ou $0,0025$, ce qui correspond à un nombre d'événements positifs prévu de 320, 64, 32 ou 8 sur un total possible de 320. Dans le tiers intermédiaire du tableau, la probabilité associée aux grappes est égale à $p/2$ avec une probabilité de $1/2$ ou à $3p/2$ avec une probabilité de $1/2$. En faisant varier p_i d'une grappe à dans le premier tiers. En faisant varier p_i d'une grappe à l'autre, une corrélation se développe entre les observations de la grappe. Cette corrélation est égale à $0,00278, 0,0051, 0,0025$ et $0,0006$ dans le tiers intermédiaire (quand on néglige les poids de l'échantillonnage) et correspond à $320, 64, 32$ ou 8 événements positifs prévus, respectivement. Pour le tiers inférieur du tableau, la probabilité est égale à $p/2$ avec une probabilité de $3/4$ ou à $5p/2$, avec une probabilité de $1/4$, ce qui correspond à une corrélation intraclasses de $0,0833, 0,0153, 0,0076$ et $0,0019$. Un poids d'échantillonnage de 1 ou de 10 est attribué au hasard dans les simulations du tableau 1, avec une probabilité de $1/2$ pour chaque observation (poids sans valeur informative). Les résultats du tableau 1 conviennent aux limites inférieure et supérieure d'un intervalle de confiance unilatéral de 95 %, idéalement, le pourcentage représentant la non-couverture dans le tableau devrait être inférieur ou égal à la valeur nominale de 5,0. Les résultats conviennent aussi aux intervalles bilatéraux de 90 %, pour lesquels, idéalement, les valeurs supérieures et inférieures du tableau devraient être $\leq 5,0$ (Jennings 1987). Pour chaque ligne du tableau, on a simulé 100 000 jeux de données au moyen du générateur de nombre aléatoire du SAS (1990, p. 631), de manière à estimer la probabilité de non-couverture des limites de l'intervalle.

Nous proposons une simple modification aux intervalles binomiaux susceptible d'en permettre l'application aux proportions estimées à partir des données d'une enquête complexe. Nous nous intéressons particulièrement au cas où un petit nombre d'événements positifs est prévu. Beaucoup d'analyses d'enquête ne fourniront pas de proportions estimées en pareil cas, car elles manqueraient de fiabilité. Si on appliquait le critère de l'erreur-type relative aux proportions de l'Enquête-ménages nationale sur la toxicomanie de 1996 (SAMHSA 1998), par exemple, la proportion estimée de femmes cocaïnomanes n'apparaîtrait pas au tableau 7. Or, nous pensons que de telles proportions véhiculent une information intéressante pourvu qu'on en précise de façon explicite l'imprécision en donnant les intervalles de confiance. On trouvera à la partie 2 les intervalles que nous proposons et ceux reposant sur la transformation par logit et la distribution de Poisson suggérés par d'autres auteurs. À la partie 3, les résultats d'une simulation comparent l'efficacité des intervalles en question. On constate que les intervalles proposés donnent de bons résultats pour ce qui est de la probabilité de couverture de la proportion réelle et de la largeur moyenne de l'intervalle. La partie 4 présente deux applications des intervalles dans une enquête majeure où l'on s'attend à un faible nombre d'événements positifs. L'article se termine par une analyse de travaux connexes débouchant sur des intervalles de confiance qui atteindront leur probabilité de couverture nominale quelle que soit la configuration des groupes au sein de la population.

2. LIMITES DES INTERVALLES PROPOSÉS ET DES AUTRES INTERVALLES DE CONFIANCE

Pour l'intervalle de confiance $1 - \alpha$ d'un échantillon de taille n , définissons d'abord la taille efficace de l'échantillon par

$$n^* = \frac{\text{var}(\hat{p})}{\hat{p}(1-\hat{p})} \quad (2.1)$$

et la taille efficace de l'échantillon ajustée selon le nombre de degrés de liberté par

$$n_{df}^* = \left(\hat{p}(1-\hat{p}) \right)^{-1} \left(\frac{\text{var}(\hat{p})}{\hat{p}(1-\hat{p})} \right)^2 \quad (2.2)$$

Les valeurs n^* et n_{df}^* sont égales à n quand $\hat{p} = 0$. Les limites proposées remplacent n par n_{df}^* , et x par $\hat{p}n_{df}^*$ dans (1.2), viz. $P_L(\hat{p}n_{df}^*, n_{df}^*)$ et $P_U(\hat{p}n_{df}^*, n_{df}^*)$. (Quand n est élevé, on peut se servir du quantile $1 - \alpha/2$ de la distribution normale au lieu de t_{n-1}^* dans (2.2).) Pour estimer l'intervalle de confiance d'une proportion dans un sous-groupe de la population, on suppose que la taille de l'échantillon n est égale à celle de l'échantillon restreint au sous-groupe.

Voici la justification heuristique de la procédure qui précède. La taille efficace de l'échantillon (2.1) est n divisée par un estimateur de l'effet du plan de sondage de l'enquête. La chose paraît raisonnable si on veut tenir compte de la variabilité supplémentaire de \hat{p} attribuable à l'échantillonnage complexe. La variabilité de l'estimateur de la variance présente elle aussi de l'importance lorsqu'on construit les intervalles de confiance. La deuxième traction de (2.2) tient compte du fait que $\text{var}(\hat{p})$ variera typiquement plus qu'un estimateur de la variance utilisé avec un échantillonnage aléatoire simple. Si d est élevé, ce facteur prend presque la valeur un et est inutile. Avec une faible valeur d et une valeur n et $\hat{p}n_{df}^*$ élevées, il serait préférable que l'intervalle proposé se rapproche de l'intervalle (1.1), qui convient en pareil cas. Puisque $F_{n,w}^w(\beta) \approx 1 + z(\beta)\sqrt{2(1/w + 1/w)}$ pour les valeurs n et w élevées (Johnson et Kotz 1970, p. 81), c'est effectivement ce qui se produit, à savoir $\hat{p} - P_L(\hat{p}n_{df}^*, n_{df}^*) \approx P_U(\hat{p}n_{df}^*, n_{df}^*) - \hat{p} \approx t_p^*(1 - \alpha/2)$. Breeze (1990) a élaboré une méthode analogue à celle que nous suggérons pour l'Enquête générale sur les ménages du Royaume-Uni. Cette méthode repose sur l'intervalle de confiance $1 - \alpha$ de l'échantillonnage aléatoire simple ($P_{0L}(x), P_{0U}(x)$) pour une variable de Poisson x , s'exprimant comme suit (Johnson et coll. 1993, p. 171):

$$P_{0L}(x) = 0.5\chi_{v_2}^2(\alpha/2) \text{ et } P_{0U}(x) = 0.5\chi_{v_2}^2(1 - \alpha/2)$$

où $v_1 = 2x$, $v_2 = 2(x + 1)$ et $\chi_{v_2}^2(\beta)$ correspond au quantile β d'une distribution χ^2 à v degrés de liberté. On suppose que l'intervalle de confiance est égal à ($P_{0L}(\hat{p}n^*)/n^*$, $P_{0U}(\hat{p}n^*)/n^*$) avec les données d'une enquête complexe. Une troisième façon de bâtir un intervalle de confiance consiste à recourir à la transformation par logit. L'intervalle de confiance de niveau $1 - \alpha$ correspond à

$$\left(\frac{1}{1 + \exp(-LLOGIT)}, \frac{1}{1 + \exp(-ULOGIT)} \right)$$

où

$$LLOGIT = \log \frac{\hat{p}}{1-\hat{p}} - t_p^*(1 - \alpha/2) \quad (2.3)$$

et

$$ULOGIT = \log \frac{\hat{p}}{1-\hat{p}} + t_p^*(1 - \alpha/2) \quad (2.4)$$

On a suggéré d'utiliser ces intervalles avec un quantile à distribution normale plutôt qu'à distribution t pour l'Enquête-ménages nationale sur la toxicomanie de 1996 (SAMHSA 1998). En dehors des conditions d'enquête complexe, quand $\hat{p} = 0$, on pourrait ajouter une petite constante au nombre d'événements et de non-événements observés, par ex. $1/2$, afin de calculer l'intervalle de

Intervalle de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête

EDWARD L. KORN et BARRY I. GRAUBARD¹

RÉSUMÉ

En dehors des enquêtes complexes, on applique fréquemment des intervalles de confiance «exacts» aux proportions obtenues par distribution binomiale au lieu d'intervalles reposant sur une normalité approximative, quand les événements positifs sont peu nombreux. Malheureusement, il est impossible d'en faire autant avec les données des enquêtes complexes, en sorte qu'on utilise des intervalles reposant sur la normalité approximative hypothétique de la proportion pondérée selon l'échantillon, même quand il existe peu d'événements positifs. Les auteurs proposent une simple modification aux intervalles binomiaux dans une situation de ce genre. Les simulations restreintes qu'ils présentent indiquent que la couverture probable de tels intervalles dépasse celle des intervalles reposant sur la normalité, des intervalles venant de la transformation par logit et des intervalles résultant de l'approximation de Poisson. Ils appliquent ensuite ces intervalles à la prévalence du virus de l'immunodéficience humaine (VIH) selon les données de la troisième Enquête nationale sur la santé et la nutrition et à la proportion de cocaïnomanes selon les données de l'Enquête sur la santé et la nutrition des personnes hispaniques.

MOTS CLÉS : Intervalle de confiance binomial; intervalle de confiance exact; transformation par logit; intervalle de confiance de Poisson.

1. INTRODUCTION

Avec les données d'une enquête complexe, l'intervalle de confiance de niveau $1 - \alpha$ typique d'une proportion d'événements positifs pour une variable de type 0-1 ressemble à

$$\hat{p} \pm t_p^* (1 - \alpha/2) [\text{var}(\hat{p})]^{1/2} \quad (1.1)$$

où \hat{p} représente l'estimateur de la proportion pondéré pour l'échantillon, où $\text{var}(\hat{p})$ est l'estimateur de la variance de \hat{p} , et où $t_p^* (1 - \alpha/2)$ correspond au $1 - \alpha/2$ quantile d'une distribution t à d degrés de liberté. On calcule l'estimateur $\text{var}(\hat{p})$ par linéarisation ou par répétition au moyen d'une méthode tenant compte du plan d'échantillonnage, notamment le fait que \hat{p} est un estimateur pondéré. Par données d'une enquête complexe, on entend les données recueillies en vertu d'un plan d'échantillonnage à plusieurs degrés, avec sélection stratifiée de grappes au premier degré. Avec un plan d'échantillonnage de ce genre, d correspond habituellement au nombre de grappes échantillonnées moins le nombre de strates (Korn et Graubard 1990). L'intervalle de confiance (1.1), que nous appellerons «intervalle linéaire» repose sur l'hypothèse que \hat{p} a une distribution proche de la normale. L'hypothèse se vérifie pour diverses asymptotes raisonnables (Krewski et Rao 1981). L'emploi du quantile t au lieu d'un quantile à distribution normale dans (1.1) s'appuie sur des preuves empiriques (Frankel 1971, ch. 7), mais il est possible de le justifier formellement à partir d'hypothèses fermes (Korn et Graubard 1990).

Avec un nombre restreint d'événements positifs cependant, \hat{p} perd sa normalité approximative (Cochran 1977, p. 58). On peut éviter l'hypothèse de la normalité en utilisant l'intervalle de confiance de Clopper et Pearson (1934) reposant sur la distribution binomiale lorsqu'on est en présence d'un échantillon aléatoire simple (ou qu'il ne s'agit pas d'une enquête complexe); Vollset (1993) donne une analyse complète des intervalles de confiance des proportions dans un tel contexte. Quand un échantillon aléatoire simple de taille n comporte x valeurs positives, l'intervalle de confiance Clopper et Pearson ($P_L(x, n)$, $P_U(x, n)$) de niveau $1 - \alpha$ peut s'exprimer comme suit (Johnson, Kotz et Kemp 1993, p. 130):

$$P_L(x, n) = \frac{v_1 F_{v_1, v_2}(\alpha/2)}{v_1 F_{v_1, v_2}(\alpha/2) + v_2} \quad (1.2)$$
$$P_U(x, n) = \frac{v_3 F_{v_3, v_4}(1 - \alpha/2)}{v_3 F_{v_3, v_4}(1 - \alpha/2) + v_4}$$

où $v_1 = 2x$, $v_2 = 2(n - x + 1)$, $v_3 = 2(x + 1)$, $v_4 = 2(n - x)$ et $F_{d_1, d_2}(\beta)$ correspond au quantile β d'une distribution F à d_1 et d_2 degrés de liberté. Pour un intervalle de confiance unilatéral, on utilise α au lieu de $\alpha/2$ dans les expressions ci-dessus. On sait qu'avec un échantillon aléatoire simple, la probabilité de couverture de tels intervalles est supérieure ou égale à leur niveau nominal, peu importe le nombre d'événements positifs. On parle parfois d'intervalles de confiance «exacts», mais nous dirons plutôt «intervalles binomiaux».

¹ Edward L. Korn, Biometric Research Branch, EPN-739, National Cancer Institute, Bethesda, MD 20892, U.S.A.; Barry I. Graubard, Biostatistics Branch, National Cancer Institute, Bethesda, MD 20892, U.S.A.

est utilisée dans la stratégie 1, on peut examiner la sensibilité de l'efficacité de X_2 pour la stratégie 1, en utilisant différentes valeurs de β , selon la méthode Prasad et Srivenkataramana (1980). Lorsqu'on utilise une valeur provisoire $\beta_0 = \beta$, la variance minimale de X_2 pour la stratégie 1 devient

$$V^{\min}(X_2 | \beta) = k(1 - f + \sqrt{\zeta^{**}}) \sigma_z^2 / 2 = M_{\beta}^2 \quad (9)$$

$$\text{où } \zeta^{**} = [1 - (1 - v^2) \delta^2] \zeta \text{ et } v = 1 - \beta / \beta_0.$$

À partir de l'équation (9), nous constatons que la stratégie 1 proposée, basée sur une valeur provisoire pour β , donne des résultats meilleurs ou pires que ceux obtenus par la stratégie de Prasad et Graham (1994) selon que $|v| < 1$ ou $|v| > 1$. De même, la stratégie 1, avec $\beta = \beta$, est supérieure ou inférieure à celle de Chotai (1974) selon que $v^2 > 0$ ou $v^2 < 0$ ($1 - 1/\delta^2$) ($1 - 1/\zeta$). Le tableau 2 qui précède indique la sensibilité E^* de l'estimateur X_2 par rapport à la stratégie de Prasad et Graham (1994) où $E^* = V_{pg}^2 / M_{\beta}^2$. À partir du tableau 2, nous constatons que la perte avec $v > 1$ risque d'être plus élevée que le gain avec $v < 1$ pour la population 1 et 3, alors que l'inverse vaut pour la population 2.

CONCLUSION

Pour l'échantillonnage sur deux cycles, il vaut mieux utiliser les données recueillies au premier cycle, pour obtenir un estimateur efficace de l'ensemble de la population au deuxième cycle. Chotai (1974) a utilisé les données recueillies au premier cycle, au stade de l'estimation, alors que Prasad et Graham (1994) les ont utilisées pour la sélection (et donc l'estimation) de l'échantillon apparié. Nous proposons ici deux stratégies. La première s'appuie sur les données recueillies au premier cycle, pour la sélection de l'échantillon apparié, comme le proposent Prasad et Graham (1994), et aussi pour la définition d'un estimateur de régression selon la méthode proposée par Chotai (1974). Il en résulte que la stratégie 1 est plus efficace que celle de Prasad et Graham (1994). La stratégie 2 utilise les valeurs du premier cycle comme variable de stratification, comme mesure de la taille pour la sélection de l'échantillon apparié au deuxième cycle et pour la formation d'un estimateur de régression utilisant la variable auxiliaire (z) obtenue au premier cycle. Intuitivement, on peut s'attendre à ce que la stratégie 2 donne de meilleurs résultats que les autres mentionnées ici; cependant, aucun

REMERCIEMENTS

L'auteur aimerait remercier l'examinateur, le rédacteur adjoint et le rédacteur en chef pour leurs précieux commentaires qui ont permis d'améliorer sensiblement la version antérieure du présent article. Ces travaux ont été réalisés grâce à une aide du FRD, en Afrique du sud.

BIBLIOGRAPHIE

- ARNAB, R. (1991). On sampling over two occasions using varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, 43(3), 282-290.
- AVADHANI, M.S., et SUKHATME, B.V. (1970). A comparison of two sampling procedures with an application to successive sampling. *Applied Statistics*, 19, 251-259.
- CHOTAI, J. (1974). A note on the Rao-Hartley-Cochran method for PPS sampling over two occasions. *Sankhyā, Series C*, 36, 173-180.
- PRASAD, N.G.N., et SRIVENKATARAMANA, T. (1980). A modification to the Horvitz-Thompson estimator under the Midzuno sampling scheme. *Biometrika*, 67, 709-711.
- PRASAD, N.G.N., et GRAHAM, J.E. (1994). Échantillonnage avec PPT en deux occasions. *Techniques d'enquête*, 20, 63-68.
- RAJ, D. (1965). On sampling over two occasions with probabilities proportional to size. *Annals of Mathematical Statistics*, 36, 327-330.
- RAO, J.N.K., HARTLEY, H.O., et COCHRAN, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society*, 24, 482-491.
- SINGH, D., et CHAUDHURI, F.S. (1986). *Theory and Analysis of Sample Survey Designs*. India: Wiley Eastern Limited, 166.
- SINGH, M.P. (1967). The relative efficiency of some two-phase sampling schemes. *Annals of Mathematical Statistics*, 38, 937-940.
- SUKHATME, P.V., et SUKHATME, B.V. (1970). *Sampling Theory of Surveys with Applications*. Ames, Iowa: Iowa State University Press, 185.
- UNITED NATIONS (1992). *Statistical Year Book*, (1988/89). New York: United Nations, 356.

superficie en blé en 1936 était supérieure ou inférieure à 200 acres. Les paramètres pour cette population sont: $N = 34$, $N_1 = 20$, $N_2 = 14$, $\delta^* = 0,7635$, $\delta = 0,3638$, $\zeta = 0,3811$, $\theta = 0,2436$. La population 2 représente la production de céréales en Amérique du sud en 1980 (y_1), 1988 (y_1) et 1989 (y_2), selon les données compilées dans l'Annuaire statistique des Nations-Unies (1988-1989). Cette population est elle aussi stratifiée en deux strates, selon que la production en 1988 est supérieure ou inférieure à 570 (milliers de tonnes métriques). Les paramètres pour cette population sont: $N = 19$, $N_1 = 7$, $N_2 = 12$, $\delta^* = -0,6939$, $\delta = 0,7666$, $\zeta = 1,1478$, $\theta = 0,3681$. Enfin, la population 3 compilée par Singh et Chaudhuri (1986) fait référence à la superficie ensemencée en blé, en hectares, en 1979-1980 (y_2) et 1978-1979 (y_1) et à la superficie totale en culture en 1978-1979 (z) dans 16 villages du district de Meerut. Les paramètres pour la population 3 sont: $N = 16$, $N_1 = 9$, $N_2 = 7$, $\delta^* = 0,7729$, $\delta = 0,1057$, $\zeta = 0,3965$, $\theta = 0,2827$. Le tableau qui suit indique l'efficacité relative des stratégies proposées 1 et 2 ainsi que de la stratégie proposée par Prasad et Graham (1994) par rapport à celle de Chotai (1974), efficacité qui est représentée respectivement par $E_1 = V^c/M_1$, $E_2 = V^c/V^{PG}$, $E_3 = V^c/M_2$ et $E_3 = V^c/V^{PG}$.

Tableau 1

Efficacité des diverses stratégies

f	Population 1			Population 2			Population 3		
	E_1	E_2	E_3	E_1	E_2	E_3	E_1	E_2	E_3
0,05	1,0463	1,1033	1,0181	1,0196	1,0850	0,8262	1,0053	1,0864	1,0030
0,10	1,0479	1,0895	1,0187	1,0202	1,0711	0,8212	1,0055	1,0711	1,0031
0,15	1,0496	1,0776	1,0194	1,0209	1,0579	0,8172	1,0057	1,0577	0,0033
0,20	1,0514	1,0683	1,0200	1,0216	1,0519	0,8123	1,0058	1,0469	1,0034
0,25	1,0533	1,0622	1,0208	1,0224	1,0490	0,8071	1,0061	1,0396	1,0035
0,30	1,0554	1,0604	1,0216	1,0232	1,0530	0,8017	1,0063	1,068	1,0036

À la lumière du tableau qui précède, nous constatons que, pour les trois populations à l'étude, la stratégie 2 donne de meilleurs résultats que les autres. Il convient également de souligner que les deux stratégies proposées donnent toutes deux de meilleurs résultats que celle de Chotai (1974) ou encore celle de Prasad et Graham (1994). Pour la population 1, $\zeta = 0,3811$, ce qui favorise grandement la stratégie de Prasad et Graham (1994) et, partant, la stratégie 1 proposée. La stratégie de Prasad et Graham (1994) et la stratégie 1 ont toutes deux avancé la stratégie de Chotai (1974). Pour la population 2, $\zeta = 1,1478$, ce qui est une valeur à la fois élevée et défavorable en regard de la stratégie de Prasad et Graham (1994) tandis que $\delta = 0,7666$ favorise bien la stratégie 1. Par conséquent, pour la population 2, la stratégie de Prasad et Graham (1994) devient moins efficace que celle de Chotai (1974), mais la stratégie 1 proposée demeure préférable. Enfin, pour la

Tableau 2

Sensibilité de l'efficacité $E^* = V^{PG}/M_\beta$

$ v $	f			Population 1			Population 2			Population 3		
	0,1	0,15	0,2	0,25	0,3		0,25	0,3		0,25	0,3	
0,0	1,028	1,029	1,030	1,031	1,032	1,033	1,027	1,028	1,029	1,031	1,032	1,033
0,2	1,027	1,027	1,028	1,029	1,031	1,032	1,024	1,027	1,028	1,029	1,031	1,032
0,4	1,023	1,024	1,027	1,026	1,027	1,028	1,017	1,020	1,021	1,023	1,026	1,027
0,6	1,017	1,018	1,019	1,019	1,020	1,021	1,010	1,011	1,011	1,011	1,011	1,011
0,8	1,010	1,010	1,010	1,010	1,011	1,011	1,000	1,000	1,000	1,000	1,000	1,000
1,0	1,000	1,000	1,000	1,000	1,000	1,000	0,989	0,988	0,988	0,988	0,987	0,987
1,2	0,989	0,988	0,988	0,988	0,988	0,987	0,976	0,976	0,975	0,974	0,973	0,972
1,4	0,883	0,880	0,877	0,875	0,871	0,869	1,234	1,241	1,249	1,257	1,266	1,278
0,0	1,002	1,002	1,004	1,003	1,003	1,003	1,219	1,227	1,233	1,241	1,249	1,258
0,2	1,002	1,002	1,002	1,002	1,003	1,002	1,180	1,186	1,191	1,197	1,204	1,211
0,4	1,002	1,002	1,002	1,002	1,002	1,002	1,125	1,128	1,133	1,137	1,141	1,146
0,6	1,002	1,002	1,002	1,002	1,002	1,001	1,063	1,065	1,067	1,068	1,070	1,073
0,8	1,001	1,001	1,001	1,001	1,001	1,001	1,000	1,000	1,000	1,000	1,000	1,000
1,0	1,000	1,000	1,000	1,000	1,000	1,000	0,999	0,999	0,999	0,999	0,999	0,999
1,2	0,999	0,999	0,999	0,999	0,999	0,999	0,976	0,976	0,975	0,974	0,973	0,972
1,4	0,998	0,997	0,998	0,998	0,998	0,998	0,939	0,938	0,936	0,935	0,933	0,931

population 3, $\zeta = 0,3965$ ce qui est assez bon pour la stratégie de Prasad et Graham (1994) mais parallèlement $\delta^* = 0,7729$ et ceci (δ^*) favorise la stratégie de Chotai (1974). En fait, la stratégie de Chotai (1974) est légèrement inférieure à celle de Prasad et Graham (1994) mais la stratégie 2 proposée demeure supérieure aux deux. Il convient toutefois de préciser que les exemples présentés ici sont assez inhabituels, en ce qu'ils représentent de faibles corrélations entre y_2 et z (dans l'exemple 1, $\delta = 0,3638$ et dans l'exemple 3, $\delta = 0,1057$) et qu'il y a une corrélation négative entre y_2 et y_1 ($\delta^* = -0,6939$) dans l'exemple 2. On s'attend à ce que les corrélations δ et δ^* soient élevées et positives. Il faudra donc mener d'autres études pour comparer le rendement des stratégies présentées au moyen de données appropriées.

Pour étudier l'effet de la déviation par rapport à la valeur optimale de $\beta = \beta_0$ lorsqu'une valeur provisoire de β

apparié s_m et l'échantillon non apparié s_n , se définissent respectivement comme suit:

$$\hat{F}_{2m}^i = \sum_{h=1}^H w_h \hat{F}_{2m}^i(h); \hat{F}_{2n}^i = \sum_{h=1}^H (y_{2i}/p_i) p_i^i \quad (9)$$

où

$$\hat{F}_{2m}^i(h) = \sum_{i=1}^{s_m} r_i^i(h) \hat{Q}_{hi}^i / (n_{1h} p_{hi} q_{hi}^* + c_h \sum_{j=1}^{s_{1h}} z_j^i(h))$$

$$(n_{1h} p_{hi}), w_h = n_{1h} / n, p_{hi} = z_j^i(h) / Z,$$

$$r_i^i(h) = y_{2i}(h) - c_h y_{1i}(h),$$

\hat{Q}_{hi}^i = somme de q_{hj}^* pour le groupe qui contient la i -ième unité de la h -ième strate créée pour la construction de l'échantillon apparié s_m par la méthode de RHC. c_h sont des constantes choisies pour réduire au minimum la variance de $\hat{F}_{2m}^i(h)$. Selon Arnab (1991), l'expression de la variance de \hat{F}_{2m}^i est définie par la formule:

$$V(\hat{F}_{2m}^i) = k_2 \sum_{h=1}^H \sum_{j=1}^h q_{hj}^i (r_{hj}^i / q_{hj}^i - R_h^i)^2 / P(h) + \sigma_2^i / n$$

où $k_2 = (n - m) / n, q_{hj}^i y_{1i}(h) / y_1(h), X_1^i(h) = \sum_{j=1}^N$ taille de la population dans la h -ième strate, $P(h) = Z_h / Z, Z = \sum_{j=1}^{N_h} z_j^i(h)$. La valeur optimale de c_h qui réduit au minimum la valeur de $V(\hat{F}_{2m}^i)$ et la valeur correspondante de $V(\hat{F}_{2m}^i)$, sont définies respectivement par

$$\text{opt } c_h = c_h(0) = \delta_{h3} = \sum_{h=1}^J q_{hj}^i \alpha_{hj} \beta_{hj} / (\sigma_{h0} \sigma_{h3})$$

et $[1 + (n - m) \theta / m] \sigma_2^i / n$, où

$$\alpha_{hj} = y_{2j}(h) / q_{hj} - X_2(h), \beta_{hj} = z_{hj} / q_{hj} - Z_h,$$

$$\sigma_2^i = \sum_{h=1}^J q_{hj}^i \alpha_{hj}^2, \sigma_2^{h0} = \sum_{h=1}^J q_{hj}^i \beta_{hj}^2, X_2(h) = \sum_{h=1}^J y_{2j}(h)$$

$$\text{et } \theta = \sum_h (1 - \delta_2^i) \sigma_2^i / \{P_h \sigma_2^i\}.$$

L'estimateur composite proposé pour X_2 , la proportion optimale de l'échantillon apparié et l'expression de la variance minimale de l'estimateur composite \hat{F}_2 sont représentés respectivement par

3. EFFICACITÉ DES STRATÉGIES PROPOSÉES

où \hat{F}_{2m}^i et \hat{F}_{2n}^i sont définis en (9), $f^* = N / (N - 1)$, $\mu_0 = 1 - \lambda_0; k, f$ et σ_2^i sont définis en (4).

$$= M_2 \text{ (disons)}$$

$$V^{\min}(\hat{F}_2^i) = k(1/\mu_0 - f) \sigma_2^i / [1 + (\lambda_0/\mu_0) \sqrt{f^*} / \sqrt{\theta}]$$

$$\text{opt } \lambda = \lambda_0 = [\theta - (1 - f) \sqrt{\theta} \sqrt{f^*}] / [\theta + f \sqrt{\theta} \sqrt{f^*} - 1]$$

$$\hat{F}_2 = \phi \hat{F}_{2m}^i + (1 - \phi) \hat{F}_{2n}^i$$

Pour établir des comparaisons, trois ensembles de données sont examinés. Le premier (que nous désignerons population 1) a été étudié par Prasad et Graham (1994); il fait référence à la superficie enssemencée en blé en 1937

(y_2) et en 1936 (y_1) ainsi qu'à la superficie en culture (z) dans un ensemble de 34 villages de l'Inde, selon des données compilées par Sukhatme et Sukhatme (1970). La

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

population 1 a été divisée en deux strates, selon que la

Théorème 1

$$V(\hat{V}_{2m}^*) = \{k/(1-k)\} \left[\sum_{i \in s_m} (y_{2i} \Delta_i / d_i^2) \tilde{p}_i / d_i^* - \hat{V}_{2m}^{*2} \right] + \{k_2/k\} \sum_{i \in s_m} \tilde{p}_i / d_i^* - \sum_{i \in s_m} \tilde{p}_i r_i^* / d_i^* \left(\tilde{p}_i / d_i^* \right)^2$$

est un estimateur sans biais de $V(\hat{V}_{2m}^*)$, lorsque β_0 est connu, $k = (N-n)/\{n(N-1)\}$ et $k_2 = (n-m)/\{m(n-1)\}$.

Théorème 2

$E_1 V_2 [\sum_{i \in s_m} \tilde{p}_i^* / d_i^*] = N(n-m)/\{nm(N-1)\} [\sigma_2^2 + \sigma_0^2 - 2\sigma_{03}]$ peut être estimé sans biais à partir de l'équation

$$\{(n-m)/n(m-1)\} \sum_{i \in s_m} (\tilde{p}_i^* / d_i^*)^2 - \sum_{i \in s_m} \tilde{p}_i^* / d_i^* \cdot {}^2 \tilde{p}_i$$

où $\tilde{p}_i^* = \tilde{p}_i \Delta_i / d_i$, $\tilde{p}_i^* = y_{2i} - z_i$, σ_2^2 , σ_0^2 et σ_{03} sont définis respectivement en (4) et (7).

À partir du théorème 2, nous constatons que

$$\hat{\sigma}_2^2 = d \sum_{i \in s_m} \left(z_i / d_i^* - \sum_{i \in s_m} z_i \tilde{p}_i / d_i^* \right) {}^2 \tilde{p}_i, \quad \hat{\sigma}_3^2 = d \sum_{i \in s_m} \left(y_{2i} / d_i^* - \sum_{i \in s_m} y_{2i} \tilde{p}_i / d_i^* \right) {}^2 \tilde{p}_i$$

et

$$\hat{\sigma}_{30}^2 = d \sum_{i \in s_m} \left(z_i / d_i^* - \sum_{i \in s_m} z_i \tilde{p}_i / d_i^* \right) \left(y_{2i} / d_i^* - \sum_{i \in s_m} y_{2i} \tilde{p}_i / d_i^* \right) \tilde{p}_i$$

sont des estimateurs sans biais de σ_2^2 , σ_3^2 et σ_{30}^2 , respectivement, où $d = m(N-1)/\{N(m-1)\}$.

Estimateur de $V^{\text{opt}}(\hat{V}_2 \lambda)$

Par conséquent, pour une valeur donnée de m ($i.e.$, λ), nous pouvons proposer un estimateur sans biais approximatif de $V^{\text{opt}}(\hat{V}_2 \lambda)$ qui corresponde à

$$V^{\text{opt}}(\hat{V}_2 \lambda) = (1/\hat{V}_m^* + 1/\hat{V}_n^*)^{-1},$$

où $\hat{V}_m^* = V(\hat{V}_{2m}^*) I \beta_0$ et $\hat{V}_n^* =$ estimateur sans biais de $V(\hat{V}_{2n}^*) = \{N-n\}/N(n-1) \sum_{i \in s_n} p_i' (y_{2i}/d_i - \hat{V}_{2n}^*)^2$.

2.2. Stratégie 2

ci-haut.

Idéalement, l'estimation de σ_2^2 devrait se faire par la combinaison optimale de $\hat{\sigma}_2^2(m)$ et $\hat{\sigma}_2^2(n)$ mais, dans le cas présent, une telle combinaison ferait intervenir des paramètres inconnus. Pour éviter cette source de complexité, l'estimateur plus simple ($\hat{\sigma}_2^2$) de σ_2^2 a été proposé

Remarque 2.2

$\hat{\sigma}_2^2(m)$ = estimateur sans biais approximatif de σ_2^2 basé sur l'échantillon apparié $s_m^* = \sum_{i \in s_m} (y_{2i} \Delta_i / d_i^*) \tilde{p}_i / d_i^* - \{ \hat{V}_{2m}^{*2} - \hat{V}_m^* \}$, $\hat{\sigma}_2^2(n)$ = estimateur sans biais approximatif de σ_2^2 basé sur l'échantillon non apparié $s_n^* = n(N-1)/\{N(n-1)\} \sum_{i \in s_n} p_i' (y_{2i} \tilde{p}_i / d_i - \hat{V}_{2n}^*)^2$; k et f correspondent aux valeurs définies dans (4).

$$\hat{\sigma}_2^2 = \lambda \hat{\sigma}_2^2(m) + (1-\lambda) \hat{\sigma}_2^2(n)$$

$$\hat{\delta} = \hat{\sigma}_{03} / (\hat{\sigma}_2^2 \hat{\sigma}_3^2)^{1/2}, \quad \hat{\zeta}^* = \hat{\sigma}_2^2 / \hat{\sigma}_2^2,$$

$$\hat{\zeta}^* = (1 - \hat{\delta}^2) \hat{\zeta}, \quad \hat{\lambda} = \sqrt{\hat{\zeta}^* / (1 + \sqrt{\hat{\zeta}^*})},$$

où

$$V^{\min}(\hat{V}_2) = k[1-f + (1-\lambda) \hat{\zeta}^* \lambda] / \hat{\sigma}_2^2,$$

à

Si l'on introduit des estimateurs appropriés de λ , ζ^* et σ_2^2 dans l'expression pour $V^{\min}(\hat{V}_2)$, nous obtenons un estimateur sans biais approximatif de $V^{\min}(\hat{V}_2)$ qui correspond

Estimateur de $V^{\min}(\hat{V}_2)$

La population est censée être composée de L strates, N_h désignant ici la taille inconnue de la h -ième strate ($h = 1, \dots, L$; $\sum_h N_h = N$), ce qui signifie que l'on peut déterminer la strate à laquelle appartient une unité dès que sa valeur est observée au premier cycle. Au premier cycle, l'échantillon initial s_1 de taille n a été sélectionné par PPTAR, en utilisant la taille normée p_i liée à la i -ième unité. Supposons que les n_h unités de s_1 , qui font partie de la h -ième strate, sont représentées par s_{1h} . Supposons également que $y_{1i}(h)$, $y_{2i}(h)$ désignent respectivement la valeur de la variable à l'étude pour la i -ième unité dans la h -ième strate, aux premier et deuxième cycles, et que $z_i(h)$ est la mesure de taille qui y correspond. Au deuxième cycle, des échantillons indépendants s_{mh} de taille m_h/n_h (en supposant que la valeur de chaque h est un nombre entier), sont sélectionnés par la méthode de RHC, en gardant la valeur de $\sum_h m_h = m$ fixe et en utilisant la taille normée $q_{hi} = [y_{1i}(h)/z_i(h)] / [\sum_{i \in s_1} [y_{1i}(h)/z_i(h)]]$ pour la i -ième unité de la h -ième strate. L'échantillon non apparié s_n^* a été choisi parmi l'ensemble de la population par la méthode de RHC avec la mesure de taille normée p_i pour la i -ième unité, comme dans la stratégie 1. Les estimateurs proposés pour V_2 , basés sur l'échantillon

mesure initiale de taille dans l'estimation. L'estimateur modifié proposé \hat{Y}_{2m}^* et l'estimateur composite de Y_2 se définissent comme suit:

$$\hat{Y}_{2m}^* = \sum_{i \in s_m} (y_{2i}^*/p_i^*) \bar{p}_i - \beta \left[\sum_{i \in s_m} (z_i^*/p_i^*) \bar{p}_i - Z \right]$$

$$\sum_{i \in s_m} (r_i^*/p_i^*) \bar{p}_i + \beta Z$$

où $z_i^* = z_i' \Delta_i' / p_i'$, $y_{2i}^* = y_{2i}' \Delta_i' / p_i'$, $r_i^* = r_i' \Delta_i' / p_i'$, $r_i' = y_{2i}' - \beta z_i'$ et β est une constante choisie avec soin de manière à réduire au minimum la variance de \hat{Y}_{2m}^* ; p_i^* , \bar{p}_i et Δ_i' correspondent aux descriptions indiquées à la section 1.3.4;

$$\hat{Y}_2 = \phi \hat{Y}_{2m}^* + (1 - \phi) \hat{Y}_{2n}$$

où \hat{Y}_{2n} est obtenu en (2).

Si $E_1(V_1)$ désigne l'espérance mathématique non con-

ditionnelle (variance) pour la sélection de l'échantillon s_1 , et $E_2(V_2)$ désigne l'espérance mathématique condition-

nelle (variance) par rapport à s_m lorsque la valeur de s est fixe, la variance de \hat{Y}_{2m}^* pour une valeur donnée de β

correspond à

$$V(\hat{Y}_{2m}^* I \beta) = E_1 V_2(\hat{Y}_{2m}^* I \beta) + V_1 E_2(\hat{Y}_{2m}^* I \beta).$$

Selon Prasad et Graham (1994), nous obtenons

$$E_1 V_2(\hat{Y}_{2m}^* I \beta) = k_1 \sigma_3^* (\beta)$$

et

$$V_1 E_2(\hat{Y}_{2m}^*) = k(1 - f) \sigma_2^2$$

où

$$k_1 = N(n - m) / \{nm(N - 1)\};$$

$$\sigma_3^* (\beta) = \sum_{i \in U} q_i (r_i' / q_i - R)^2$$

$$= \sigma_2^2 + \beta^2 \sigma_0^2 - 2\beta \sigma_0 \sigma_3 \delta;$$

$$R = \sum_{i \in U} R_i = Y_2 - \beta Z, \delta = \sigma_{03} / (\sigma_0 \sigma_3),$$

$$\sigma_2^2 = \sum_{i \in U} q_i (z_i' / q_i - Z)^2,$$

$$\sigma_{03} = \sum_{i \in U} q_i (y_{2i}' / q_i - Y_2)(z_i' / q_i - Z)$$

(7)

σ_2^2 , k et σ_3^* , q_i sont obtenus respectivement de (4) et (6). La valeur optimale de β qui réduit au minimum la valeur de $V(\hat{Y}_{2m}^* I \beta)$ est définie par la formule $\text{opt } \beta = \delta \sigma_3 / \sigma_0$. Si l'on introduit la valeur optimale de $\beta = \beta_0$ dans l'expression de $V(\hat{Y}_{2m}^* I \beta)$, on obtient la valeur optimale de

$$V(\hat{Y}_{2m}^* I \beta) = V(\hat{Y}_{2m}^* I \beta_0) = k[(1 - f) + (1 - \lambda) \zeta^* / \lambda] \sigma_2^2$$

où $\zeta^* = (1 - \delta^2) \zeta$; k , f et ζ sont définis respectivement

dans (4) et (6).

On obtient la variance optimale de \hat{Y}_2 pour une valeur donnée de λ en réduisant au minimum la variance de \hat{Y}_2 par rapport à ϕ lorsque $\beta = \beta_0$, et la variance optimale est

obtenue par la formule

$$V^{\text{opt}}(\hat{Y}_2 I \lambda) = [1 / V(\hat{Y}_{2m}^* I \beta_0) + 1 / (\hat{Y}_{2n}^*)]^{-1}$$

$$= [1 / \{k(1 - f) + (1 - \lambda) \zeta^* / \lambda\} + 1 / \{k(1 - f) + (1 - \lambda) \zeta^* / \lambda\}]^{-1} \sigma_2^2$$

Enfin, en réduisant au minimum la valeur de $V^{\text{opt}}(\hat{Y}_2 I \lambda)$ par rapport à λ , la proportion optimale de l'échantillon apparié et la variance minimum de \hat{Y}_2 se définissent respectivement comme suit:

$$\text{opt } \lambda = \lambda_0 = \sqrt{\zeta^*} / (1 + \sqrt{\zeta^*})$$

et

$$V^{\text{min}}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta^*}) \sigma_2^2 / 2 = M_1 \text{ (disons)} \quad (8)$$

Remarque 2.1

L'estimateur \hat{Y}_{2m}^* , décrit en (1) peut être utilisé en pratique lorsque la valeur optimale de $\beta = \beta_0$ est connue ou que l'on possède une bonne indication de la valeur de β_0 à partir d'enquêtes précédentes. Si, au lieu de l'estimateur de régression \hat{Y}_{2m}^* , décrit précédemment, on utilise l'estimateur de différence $\hat{Y}_{2m}^{**} = \sum_{i \in s_m} (y_{2i}' / p_i') \bar{p}_i - [\sum_{i \in s_m} (z_i' / p_i') \bar{p}_i - Z]$ basé sur l'échantillon apparié, l'expression de la variance minimum de \hat{Y}_2 devient alors:

$$V^{\text{min}}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta}) \sigma_2^2 / 2 = \tilde{M}_1 \text{ (disons)}$$

avec

$$\tilde{\zeta} = (1 + \tau^2 - 2\tau\delta) \zeta, \tau = \sigma_0 / \sigma_3.$$

2.1.1 Estimation de la variance

Pour obtenir des estimateurs sans biais approximatifs de $V^{\text{opt}}(\hat{Y}_2)$, nous énonçons d'abord les théorèmes non démontrés qui suivent:

Remarque 1.1

Si l'on se rapporte aux stratégies décrites à la section 1.3, nous constatons que le plan de Avadhani et Sukhatne (1970) ne requiert pas d'information sur les mesures de taille pour l'ensemble de la base de sondage et qu'il est donc moins exigeant que les autres. Chotai (1974) a utilisé les mesures de taille initiales p_i pour la sélection, mais les valeurs obtenues lors de la première enquête y_{1i} , $i \in s_1$, ont été utilisées pour l'estimation seulement. L'utilisation de cette information supplémentaire p_i pour la sélection de l'échantillon initial s_1 rend la stratégie de Chotai (1974) plus efficace que celle de Avadhani et Sukhatne (1970). Cependant, pour utiliser l'estimateur optimal \hat{Y}_2 avec la stratégie de Avadhani et Sukhatne (1970), il faut estimer ϕ , le seul paramètre inconnu. Par contre, dans la stratégie de Chotai (1974), les deux paramètres ϕ et γ doivent être estimés pour pouvoir utiliser l'estimateur optimal \hat{Y}_2 . Prasad et Graham (1994) ont utilisé ces deux variables pour la sélection de l'échantillon apparié (donc automatiquement aussi pour l'estimation) et ils ont démontré, de façon empirique, que leur stratégie donne de meilleurs résultats que celle de Chotai (1974). Pour une plus grande efficacité, on peut également utiliser la stratégie de Prasad et Graham (1994) dans la pratique, car \hat{Y}_2 ne suppose qu'un seul paramètre inconnu, ϕ . Il est à noter que Arnab (1991) a été le premier à introduire le principe de la stratification en utilisant y_{1i} , $i \in s_1$, comme variable de stratification. Cette technique doit toujours être utilisée en pratique, chaque fois que l'information nécessaire est disponible, en particulier pour la sélection de larges unités qui présentent de grandes différences au niveau de la taille, comme celles examinées dans les exemples numériques indiqués à la section 3. On s'attend à ce que la stratégie de Arnab (1991) soit plus efficace que les stratégies précédentes, car elle utilise les valeurs du premier cycle à la fois pour la stratification et l'estimation. L'estimateur optimal \hat{Y}_2 contient toutefois plusieurs paramètres inconnus (pour plus de détails, voir Arnab 1991) qui peuvent nuire à l'application de la stratégie, en particulier lorsque la taille de l'échantillon est insuffisante.

2. STRATÉGIES PROPOSÉES

Nous examinons ici deux stratégies d'échantillonnage qui sont des modifications des stratégies proposées respectivement par Prasad et Graham (1994) et par Arnab (1991).

2.1 Stratégie 1

Le plan d'échantillonnage pour cette stratégie est le même que celui examiné par Prasad et Graham (1994), et qui est décrit à la section 1.3.4. Ici, seul l'estimateur basé sur l'échantillon apparié s_m a été modifié en introduisant la

méthode de sélection avec probabilité proportionnelle à la taille avec remise (PPTAR), en utilisant la mesure de taille normée $p_i = z_i/Z$ pour la i -ième unité. À partir des valeurs pré-établies y_{1i} ($i \in s_1$) obtenues sur la base de certains critères, les n unités de l'échantillon sont réparties en un nombre approprié de L strates. Supposons que s_{1h} représente l'échantillon de taille n_h , dans la h -ième strate ($s_1 = U_h s_{1h}$ et $\sum_h n_h = n$). Nous supposons ici que n est suffisamment grand pour s'assurer que n_h est positif pour chaque h dans la pratique. Au deuxième cycle, les sous-échantillons s_{mh} de taille $m_h (= v_h n_h, v_h$ est une fraction prédéterminée et m_h est censé être un nombre entier) sont sélectionnés de façon indépendante de s_{1h} , selon des plans d'échantillonnage appropriés utilisant y_{1i} , $i \in s_1$, pour la construction des échantillons appariés s_{mh} . L'échantillon non apparié s_u est sélectionné par PPTAR, à partir de l'ensemble de la population U , en utilisant z'_i comme mesure de taille.

1.3.4 Prasad et Graham (1994)

Ici, l'échantillon initial s_1 est sélectionné par un plan d'échantillonnage de RHC similaire à la méthode de Chotai (1974), en utilisant une mesure de taille normée $p_i = z_i/Z$ pour la i -ième unité. L'échantillon apparié s_m est prélevé de s_1 par la méthode de RHC en utilisant $p_i^* = (y_{1i} \Delta_i / p_i) / \sum_{i \in s_1} (y_{1i} \Delta_i / p_i)$ pour la i -ième unité, $i \in s_1$, où Δ_i est la somme des valeurs de p_j pour le groupe qui renferme la i -ième unité obtenue en sélectionnant s_1 par le plan d'échantillonnage de RHC. L'échantillon non apparié, s_u , a été prélevé de l'ensemble de la population U par un plan d'échantillonnage de RHC similaire à celui proposé par Chotai (1974). Ici encore, on présume que N/n , n/m et N/u sont des nombres entiers. Prasad et Graham (1994) ont proposé l'estimateur composite suivant pour Y_2 :

$$\hat{Y}_2 = \phi \hat{Y}_{2m} + (1 - \phi) \hat{Y}_{2u}$$

où $\hat{Y}_{2m} = \sum_{i \in s_m} (y_{2i} / p_i^*) \hat{p}_i$; $\hat{Y}_{2u} = \sum_{i \in s_u} (y_{2i} / p_i) \hat{p}_i$; $y_{2i}^* = y_{2i} \Delta_i / p_i$; $\hat{p}_i^* = \hat{p}_i (p_i^*) =$ total des valeurs de p_i^* (p_i^*) associées aux unités qui appartiennent au groupe aléatoire duquel la i -ième unité a été sélectionnée dans s_m (s_u). L'expression pour la variance minimum de \hat{Y}_2 est représentée par:

$$V_{\min}(\hat{Y}_2) = k(1 - f + \sqrt{\zeta}) \sigma_2^2 / 2 = V_{pg} \text{ (disons)} \quad (5)$$

$$\zeta = \sigma_2^2 / \sigma_2^2, \sigma_2^2 = \sum_{i \in U} q_i (y_{2i} / q_i - X_2)^2, q_i = y_{1i} / X_1; \quad (6)$$

k, f, σ_2^2 et X_1 sont définis en (4).

Dans l'expression de Prasad et Graham (1994) de $V_{\min}(\hat{Y}_2)$, le diviseur 2 a été omis et il s'agit de toute évidence d'une erreur typographique.

1.2 Méthode d'estimation

À partir des données $y_{1i}, i \in s_1$, et $y_{2i}, i \in s_m$, obtenues de l'échantillon initial s_1 , et de l'échantillon apparié s_m , un estimateur sans biais X_2^{zm} pour X_2 – la population totale au deuxième cycle – est obtenu en utilisant $y_{1i}, i \in s_1$ comme étant de l'information auxiliaire. Par conséquent, X_2^{zm} désigne habituellement un estimateur de la différence, un estimateur par quotient ou un estimateur de régression. À partir de l'échantillon non apparié s_n , un estimateur sans biais X_2^{zn} est également construit pour X_2 . Enfin, un estimateur composite – une combinaison de X_2^{zm} et X_2^{zn} – est obtenu en utilisant un poids approprié de ϕ ($0 \leq \phi \leq 1$)

$$X_2 = \phi X_2^{zm} + (1 - \phi) X_2^{zn}. \quad (1)$$

On obtient la valeur optimale de $\phi = \phi(\lambda)$ en réduisant au minimum la valeur de $V(X_2)$, qui désigne la variance de X_2 par rapport à ϕ pour une valeur donnée de $m(i.e., \lambda)$. Les expressions pour $\phi(\lambda)$ et $V(X_2 I \lambda)$, la variance de X_2 lorsque $\phi = \phi(\lambda)$, sont calculées comme suit, lorsque X_2^{zm} et X_2^{zn} sont indépendantes:

$$\phi(\lambda) = (1/V^{zm}) [1/V^{zm} + 1/V^{zn}]^{-1},$$

$$V(X_2 I \lambda) = [1/V^{zm} + 1/V^{zn}]^{-1},$$

où V^{zm} et V^{zn} sont les variances de X_2^{zm} et X_2^{zn} respectivement. La proportion optimale de l'échantillon apparié $\lambda = \lambda_0$ est déterminée en réduisant au minimum la valeur de $V(X_2 I \lambda)$ par rapport à λ . Enfin, lorsqu'on introduit $\lambda = \lambda_0$ dans l'expression $V(X_2 I \lambda)$, on obtient la variance minimale de X_2 , représentée par $V^{min}(X_2) = V(X_2 I \lambda_0)$. Notre objectif est de trouver une stratégie appropriée, qui soit une combinaison de $P = (P_1, P^m, P^n)$ et X_2 et qui permette de maintenir la valeur de $V^{min}(X_2)$ à un minimum.

1.3 Quelques stratégies d'échantillonnage

1.3.1 Avadhani et Sukhatme (1970)

Au premier cycle, l'échantillon initial s_1 de taille n a été sélectionné par échantillonnage aléatoire simple sans remise (EASSR), en présupposant qu'aucune information auxiliaire n'était disponible avant l'enquête. Au deuxième cycle, l'échantillon apparié s_m de taille m a été prélevé de s_1 par la méthode de Rao, Hartley et Cochran (ou méthode de RHC 1962) en utilisant y_{1i} comme mesure de la taille de la i -ième unité $i \in s_1$, en supposant que y_{1i} est positif. Selon le plan d'échantillonnage de RHC, les n unités sélectionnées de s_1 sont réparties de façon aléatoire en m groupes, chacun de taille n/m que l'on présuppose être un nombre entier. Dans chaque groupe sélectionné, une unité est choisie de façon indépendante, avec une probabilité proportionnelle à la mesure de taille. Par conséquent, si la i -ième unité, U_i , appartient au j -ième groupe G_j ($j = 1, \dots, m$), alors la probabilité de sélectionner U_i est égale à $q_i^*(i \in s_1) = y_{1i} / \sum_{i \in s_1} y_{1i}$. L'échantillon non apparié s_n a été prélevé de U/s_1 par EASSR.

1.3.2 Chotai (1974)

Au premier cycle, l'échantillon initial s_1 de taille n a été sélectionné par le plan d'échantillonnage de RHC décrit précédemment (en présupposant que N/n est un nombre entier), avec une probabilité proportionnelle à z_i , la mesure de la taille de la i -ième unité que l'on présuppose être positive et connue pour chaque $i \in U$. Supposons que $\Delta_j = \sum_{i \in G_j} P_i^k$, la somme des valeurs de P_i^k ($k = z_i/Z, Z = \sum_{i \in U} z_i$) qui appartiennent au groupe aléatoire G_j ($j = 1, \dots, n$) formé en sélectionnant l'échantillon s_1 par la méthode de RHC. L'échantillon apparié s_m a été prélevé de s_1 par le plan de RHC, en utilisant la mesure de taille normée Δ_j pour la i -ième unité $i \in s_1$ ($\sum_{i \in s_1} \Delta_j = 1$), en supposant que n/m est un nombre entier. L'échantillon non apparié, s_n a été sélectionné selon le plan d'échantillonnage de RHC, avec la mesure de taille normée P_i^j pour la i -ième unité, en présupposant la encore que N/n est un entier. Supposons que P_i^j (P_i^j) = total des valeurs de Δ_j (P_i^j) associées aux unités qui appartiennent au groupe aléatoire duquel la i -ième unité a été sélectionnée dans s_m (s_n) par la méthode d'échantillonnage de RHC avec $\sum_{i \in s_m} P_i^j = 1$ ($\sum_{i \in s_n} P_i^j = 1$).

L'estimateur composite de X_2 est représenté par

$$X_2 = \phi X_2^{zm} + (1 - \phi) X_2^{zn}$$

où

$$X_2^{zm} = \sum_{i \in s_m} (y_{2i}/P_i^j) P_i^j - \left[\sum_{i \in s_m} (y_{1i}/P_i^j) P_i^j - \sum_{i \in s_1} (y_{1i}/P_i^j) \Delta_j \right];$$

$$X_2^{zn} = \sum_{i \in s_n} (y_{2i}/P_i^j) P_i^j \quad (2)$$

où γ est une constante choisie avec soin de manière à réduire au minimum la variance de X_2^{zm} . Chotai (1974) a défini l'expression du minimum de la variance de X_2 comme étant

$$V^{min}(X_2) = k [1 - f + \sqrt{(1 - \delta^*)}] \sigma_2^2 / 2 = V^c \text{ (disons)} \quad (3)$$

où

$$k = N / \{n(N - 1)\}, f = n/N,$$

$$\sigma_2^2 = \sum_{i \in U} P_i^j (y_{2i}^2 / P_i^j - X_2^j)^2, j = 1, 2$$

$$X_2^j = \sum_{i \in U} y_{2i}^j, j = 1, 2$$

$$\delta^* = \sum_{i \in U} P_i^j (y_{2i} / P_i^j - X_2^j) (y_{1i} / P_i^j - X_1^j) / (\sigma_1 \sigma_2).$$

(4)

Arnab (1991) a proposé plusieurs stratégies selon lesquelles l'échantillon initial s_1 est construit par la

Echantillonnage en deux cycles: Estimation de la population totale

RAGHUNATH ARNAB¹

RÉSUMÉ

Deux stratégies d'échantillonnage sont proposées pour estimer les chiffres d'une population finie au cycle le plus récent, à partir des échantillons prélevés sur deux cycles, au moyen de plans d'échantillonnage à probabilité variable. Nous utilisons les données recueillies au premier cycle sur une variable de la taille et comme mesure de la taille et comme variable de stratification pour la sélection de l'échantillon apparié, au deuxième cycle. L'efficacité relative des stratégies proposées est comparée ici à d'autres méthodes appropriées.

MOTS CLÉS: Estimateur composite; échantillon apparié; plan d'échantillonnage; stratégie d'échantillonnage; plan d'échantillonnage à probabilité variable.

1. INTRODUCTION

Il arrive très souvent qu'une même population fasse l'objet d'une enquête menée à intervalles réguliers, dans le but d'estimer les mêmes caractéristiques de la population en regard des changements qui surviennent dans le temps. Ainsi, un grand nombre de pays recueillent des données sur une base annuelle ou trimestrielle pour estimer, par exemple, le nombre total de personnes en chômage, de personnes infectées par le VIH, d'immigrants, etc. Dans le présent article, nous examinons une population finie $U = (U_1, \dots, U_N, \dots, U_N)$ composée de N unités identifiables, qui est censée être échantillonnée à deux reprises pour estimer la population totale d'une variable à l'étude au deuxième cycle (actuellement). Dans le cas d'un échantillonnage successif, les données recueillies au cycle précédent (le premier) sont utilisées pour élaborer une stratégie efficace qui tienne compte des coûts et qui produise un estimateur efficace de l'ensemble de la population au cycle présent. De nombreux auteurs ont traité de ces méthodes. Singh (1967), de même que Avadhani et Sukhatme (1970), ont utilisé l'information recueillie au premier cycle comme mesure de la taille, pour sélectionner l'échantillon apparié au deuxième cycle; de son côté, Arnab (1991) a utilisé cette information à la fois comme variable de stratification et mesure de la taille pour la sélection de l'échantillon au deuxième cycle. Plus récemment, Prasad et Graham (1994) ont modifié les stratégies d'échantillonnage de Raj (1965) et de Chotai (1974), en utilisant l'information obtenue au premier cycle comme mesure de la taille pour la sélection de l'échantillon apparié au deuxième cycle; ils ont constaté, de façon empirique, qu'une de leurs stratégies donnait de meilleurs résultats que celle de Chotai (1974). Dans cet article, deux stratégies différentes sont proposées. L'une d'elle utilise l'information recueillie au premier cycle comme mesure de la taille, alors que l'autre utilise cette information à la fois comme mesure de la taille et comme

variable de stratification pour la sélection de l'échantillon apparié au deuxième cycle. Nous démontrons ici qu'une des stratégies proposées est meilleure que celle de Prasad et Graham (1994); il nous est toutefois impossible de tirer de conclusion théorique définie en ce qui concerne la deuxième stratégie. Des données empiriques montrent cependant que cette dernière stratégie est plus efficace que celle décrite par Prasad et Graham (1994) et aussi que la première stratégie proposée, du fait qu'elle utilise les valeurs du premier cycle pour tous les stades possibles, c'est-à-dire stratification, estimation et sélection de l'échantillon apparié au deuxième cycle.

Les méthodes générales de sélection et d'estimation aux deux cycles sont décrites ci-après.

1.1 Plans d'échantillonnage

Lors du premier cycle d'enquête, un échantillon s_1 , de taille n , est choisi au moyen d'un plan d'échantillonnage approprié, disons P_1 , et les données y_{1i} , $i \in s_1$, sont obtenues, où y_{1i} (y_{2i}) est la valeur de la variable aléatoire à l'étude, pour la i -ième unité au premier (deuxième) cycle. Au deuxième cycle, un échantillon apparié (sous-échantillon) s_m de taille $m (= n\lambda)$, est censé être un nombre entier, $0 \leq \lambda \leq 1$ est choisi à partir de s_1 , par un plan d'échantillonnage approprié P_m , et est complété par un échantillon non apparié s_n de taille n ($= n\mu = n - m$, $\mu = 1 - \lambda$), lequel est choisi, soit parmi l'ensemble de la population U , soit parmi U/s_1 , la série d'unités qui n'ont pas été sélectionnées au premier cycle par le plan d'échantillonnage P_n ; l'information y_{2i} ($i \in s_m$, $i \in s_n$) au deuxième cycle est ainsi obtenue. Bien sûr, on s'attend à ce que le coût de l'enquête pour les unités de l'échantillon apparié soit beaucoup moins élevé que celui pour les unités non appariées; cependant, pour simplifier l'étude, nous présumons ici que le coût d'enquête est le même pour toutes les unités, au deuxième cycle.

¹ Raghunath Arnab, Department of Statistics, University of Durban-Westville, Private Bag-X54001, Durban - 4000, South Africa.

- BAKKER, B.F.M., et WINKELS, J.W. (1998). Why integration of household surveys? – Why POLS?. *Netherlands Official Statistics*, 13, 5-7.
- BARR, R.S., et TURNER, J.S. (1980). Merging the 1977 Statistics of Income and the March 1978 Current Population Survey. Technical report, U.S. Department of the Treasury, Office of Tax Analysis.
- BETHLEHEM, J.G., et KELLER, W.J. (1987). Linear weighting of sample survey data. *Journal of Official Statistics*, 3, 141-153.
- COPELAND, K.R., PEITZMEIER, F.K., et HOY, C.E. (1987). Méthode alternative pour ajuster les estimations de la Current Population Survey aux chiffres de population. *Techniques d'enquête*, 13, 183-191.
- DEVILLE, J.C., et SÄRNDA, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.
- DEVILLE, J.C., SÄRNDA, C.-E., et SAVTORY, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, 88, 1013-1020.
- FELLEGI, I.P., et SUNTER, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.
- HOFMANS, M.G. (1998). Innovative weighting in POLS. Making use of core questions. *Netherlands Official Statistics*, 13, 12-15.
- LEMAÎTRE, G., et DUFOUR, J. (1987). Une méthode intégrée de pondération des personnes et des familles. *Techniques d'enquête*, 13, 211-220.
- MADDALA, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- OH, H.L., et SCHEUREN, F. (1987). Variante de la méthode itérative du quotient. *Techniques d'enquête*, 13, 221-232.
- PAASS, G. (1986). Statistical match: Evaluation of existing procedures and improvements by using additional information. Dans *Microanalytic Simulation Models to Support Social and Financial Policy*. Amsterdam: Elsevier Science.
- RAGHUNATHAN, T.E., et GRIZZLE, J.E. (1995). A split questionnaire survey design. *Journal of the American Statistical Association*, 90, 54-63.
- RENSSEN, R.H., et NIEUWENBROEK, N.J. (1997). Aligning estimates for common variables in two or more sample surveys. *Journal of the American Statistical Association*, 92, 368-374.
- RODGERS, W.L. (1984). An evaluation of statistical matching. *Journal of Business & Economic Statistics*, 2, 91-102.
- RUBIN, D.B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4, 87-94.
- SÄRNDA, C.-E., SWENSSON, B., et WRETMAN, J.H. (1992). *Model Assisted Survey Sampling*. New York: Springer Verlag.
- SEARLE, S.R. (1971). *Linear Models*. New York: John Wiley & Sons.
- SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley & Sons.
- SINGH, A.C., MANTTEL, H.J., KINACK, M.D., et ROWE, R. (1993). Appariement statistique: l'utilisation d'information supplémentaire comme solution de remplacement à l'hypothèse d'indépendance conditionnelle. *Techniques d'enquête*, 19, 67-89.
- TUJINEN VAN, H.K. (1995). Social indicators, social surveys and integration of social statistics. *Statistical Journal of the United Nations ECE*, 12, 379-394.
- WINKELS, J.W., et EVERAERS, P.C.J. (1998). Design of an integrated survey in the Netherlands. The case POLS. *Netherlands Official Statistics*, 13, 6-11.
- ZIESCHANG, K.D. (1990). A generalized least squares weighting system for the consumer expenditure survey. *Journal of the American Statistical Association*, 85, 986-1001.

À l'aide de méthodes d'appariement statistique, et en utilisant X et U comme variables communes, nous pouvons également obtenir une estimation dans le cas d'un tableau synthétique à double entrée composé de X et Z . Finalement, l'échantillon de petite taille a été pondéré d'après deux ensembles de variables de contrôle différents. Le premier ensemble de variables de contrôle correspondait aux totaux de population de X et Z , tandis que le deuxième ensemble de variables de contrôle correspondait au tableau synthétique à double entrée. L'utilisation du premier ensemble de variables de contrôle s'apparente fortement à la stratification double incomplète. Le cadre théorique qui est nécessaire pour créer la deuxième méthode de pondération a été présenté tout au long du présent article. À l'aide des deux méthodes de pondération, on peut estimer le tableau YZ (on présuppose tacitement que X et Z sont des variables nominales). Les dénombrements marginaux du tableau YZ relatifs à la première méthode de pondération correspondent, par définition des équations de calage, aux totaux estimés de population de X (qui est fondée sur le premier grand échantillon) et de Z (qui est fondée sur le deuxième grand échantillon). Nous avons montré que cette propriété de cohérence est confirmée dans le cas de la deuxième méthode de pondération. Nous avons également effectué une étude numérique pour évaluer le rendement des méthodes de pondération en ce qui a trait aux dénombrements de cellule. Nous avons constaté à cet égard que les deux méthodes de pondération donnent des tableaux à double entrée qui sont presque exempts de biais de plan d'échantillonnage. Les variances simulées (de plan d'échantillonnage) relatives à la deuxième méthode de pondération semblaient être plus faibles que les variances (de plan d'échantillonnage) correspondantes relatives à la première méthode de pondération, en ce qui a trait à tous les dénombrements de cellule estimés. En principe, on a présupposé que les variables X et Z étaient des variables nominales; cependant, nous avons précisé que les idées qui sont présentées peuvent être appliquées également dans le cas de variables X et Z continues, ou dans le cas d'une variable X continue et d'une variable Z nominale.

REMERCIEMENTS

L'auteur remercie Peter Kooiman, Nico Nieuwenbroek, et Ger Slootbeek pour leur lecture attentive du présent article et leurs judicieuses observations. L'auteur remercie également deux examinateurs anonymes et un rédacteur associé pour leur précieux conseils visant à améliorer l'article. Les opinions exprimées dans celui-ci sont celles de l'auteur et ne reflètent pas nécessairement la politique de Statistics Netherlands.

BIBLIOGRAPHIE

ALEXANDER, C.H. (1987). Une classe de méthodes utilisant des chiffres de population dans la pondération des ménages. *Techniques d'enquête*, 13, 193-209.

si C et Z présentent entre elles une corrélation parfaite, le côté gauche de (11) se réduit à $\sum_{k=1}^n v_{1k} y_{1k} z_{1k}$, c'est-à-dire que notre tableau estimé à double entrée correspond à un tableau estimé à double entrée pondéré fondé sur le premier échantillon, comme si les valeurs réelles de Z étaient imputées dans cet échantillon. De manière analogue, si C et X présentent entre elles une corrélation parfaite, (12) se réduit à $\sum_{k=1}^n v_{2k} y_{2k} z_{2k}$.

Un cas particulier important qui doit être pris en considération est celui où c est une variable nominale. Dans ce cas, les égalités suivantes sont vérifiées:

$$\sum_{k=1}^n v_{1k} (c_k, c'_k) = \sum_{k=1}^n v_{2k} (c_k, c'_k) = \text{diag} \begin{pmatrix} t_x \\ t_y \\ t_n \end{pmatrix},$$

ainsi (11) et (12) coïncident. En outre, nous avons, dans le cas d'une variable c nominale:

$$\sum_{k=1}^n v_{1k} c_k z'_k = \sum_{k=1}^n v_{2k} c_k z'_k$$

$$\sum_{k=1}^n v_{2k} y'_k c'_k = \sum_{k=1}^n v_{1k} y'_k c'_k$$

et

Evidemment, si c est nominale, il suffit de créer un échantillon synthétique qui est fondé soit sur le premier échantillon synthétique, soit sur le deuxième. Dans un cas comme dans l'autre, les estimations à pondération pour les tableaux CZ , CY et YZ peuvent être reconstituées. Enfin, nous signalons que les valeurs imputées dans tous les échantillons synthétiques peuvent être irréalistes. Comme nous l'avons mentionné dans la section 2.4, les prédictions calculées peuvent être remplacées par des valeurs réelles d'après un quelconque algorithme.

4. RÉCAPITULATION

Dans le présent article nous avons décrit une méthode de pondération qui permet de combiner des informations provenant de différentes enquêtes sur échantillon. La caractéristique de ces enquêtes est un ensemble de variables communes (voir figure 1). Nous prétendons que ces échantillons devraient être pondérés selon une structure séquentielle. Dans un premier temps, nous avons pondéré les deux échantillons de grande taille en utilisant les valeurs de X comme variables de contrôle. D'après ces échantillons pondérés, nous avons pu obtenir une estimation combinée du total de population de U . Ensuite, les deux échantillons de grande taille ont été pondérés à nouveau en utilisant simultanément X et U comme variables de contrôle. Cela a donné une estimation du total de population de X et Z .

Les moyennes observées au cours des 500 simulations sont presque identiques dans le cas des quatre types d'estimateur, comme le montrent ces tableaux. Il est à noter que les dénombrements de cellule indiqués sont arrondis. Nous avons également calculé le tableau YZ réel à partir de la population finie. Les dénombrements réels correspondent exactement aux moyennes qui sont indiquées dans le tableau 5 (ou 6). Dans le cas de cette simulation, nous en arrivons à la conclusion que tous les estimateurs comportent un biais très faible.

Les variances observées au cours des 500 simulations sont indiquées entre parenthèses. Les variances relatives aux dénombrements marginaux estimés des tableaux 5 et 7 coïncident parce que ces estimations sont fondées sur le même estimateur. Pour la même raison, les variances relatives aux dénombrements marginaux estimés des tableaux 6 et 8 coïncident également. Il est à noter que les variances relatives aux dénombrements marginaux estimés des tableaux 5 et 7, en raison du plus grand ensemble de variables communes. Cependant, dans le cas de la plupart des dénombrements marginaux estimés, cette réduction de la variance peut être considérée comme négligeable.

Les tableaux 5 et 6 donnent des variances identiques en ce qui a trait à tous les dénombrements estimés. Les variances dans le cas de la plupart des dénombrements estimés du tableau 7 sont clairement plus faibles que celles indiquées dans les tableaux 5 et 6. Dans le tableau 8, cette réduction de la variance est encore plus importante. Dans le cas de cet exemple, nous en arrivons à la conclusion que l'utilisation de l'ensemble plus grand de variables communes avec la première méthode de pondération réduit légèrement les taux de variance relatifs aux dénombrements marginaux estimés, mais n'a pas d'incidence sur les taux de variance des dénombrements de cellule estimés. Evidemment, l'utilisation de l'ensemble plus grand de variables communes avec la deuxième méthode de pondération réduit également légèrement les taux de variance des dénombrements de cellule marginaux. Enfin, compte tenu de l'ensemble de variables communes, la méthode de pondération fondée sur un appariement synthétique donne des variances plus faibles dans le cas des dénombrements de cellule estimés que la méthode fondée sur la stratification double incomplète.

3.4 Imputation de valeurs tirées d'un des grands échantillons dans l'autre grand échantillon

Au moyen de deux échantillons de grande taille et du petit échantillon, on peut créer un échantillon synthétique dans lequel les valeurs réelles de X et les valeurs prédites de Z , ou les valeurs réelles de X et les valeurs prédites de Z , sont enregistrées simultanément. Nous définissons des prédictions pour les valeurs de X et Z de manière analogue à (7), à savoir

$$\hat{y}_k = \hat{B}'c_k + \hat{\beta}_2'(z_k - \hat{A}'c_k), k = 1, \dots, n_2, \quad (9)$$

$$\hat{z}_k = \hat{A}'c_k + \hat{\alpha}_2'(y_k - \hat{B}'c_k), k = 1, \dots, n_1, \quad (10)$$

où

$$\hat{\beta}_2 = \left[\sum_{k=1}^{n_2} v_{2k}(z_k - \hat{A}'c_k)'(z_k - \hat{A}'c_k) \right]^{-1} \times \left[\sum_{k=1}^{n_2} v_{2k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)' \right]$$

$$\hat{\alpha}_2 = \left[\sum_{k=1}^{n_1} v_{1k}(y_k - \hat{B}'c_k)'(y_k - \hat{B}'c_k) \right]^{-1} \times \left[\sum_{k=1}^{n_1} w_{3k}(y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)' \right]$$

et

$$\sum_{k=1}^{n_1} v_{1k} y_k z_k' = \hat{B}' \sum_{k=1}^{n_1} v_{1k} c_k c_k' \hat{A} + \sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)' \quad (11)$$

et

$$\sum_{k=1}^{n_2} v_{2k} \hat{y}_k z_k' = \hat{B}' \sum_{k=1}^{n_2} v_{2k} c_k c_k' \hat{A} +$$

Dans le cas de chaque (c_k, y_k) , les valeurs de Z peuvent être imputées dans le premier grand échantillon au moyen de (10), $k = 1, \dots, n_1$, et de façon analogue, dans le cas de chaque (c_k, z_k) , les valeurs de X peuvent être imputées dans le deuxième grand échantillon au moyen de (9), $k = 1, \dots, n_2$. D'après ces valeurs imputées, nous pouvons définir les estimations suivantes pour le tableau à double entrée composé de X et Z :

$$\sum_{k=1}^{n_3} w_{3k} (y_k - \hat{B}'c_k)(z_k - \hat{A}'c_k)' \quad (12)$$

Une des estimations est fondée sur le premier échantillon synthétique, l'autre sur le deuxième échantillon synthétique. En réunissant les échantillons synthétiques, on obtient un échantillon synthétique combiné de taille $n_1 + n_2$, à partir duquel on peut établir une estimation combinée pour le tableau à double entrée. Cette estimation combinée montre une ressemblance plus étroite avec (8). Il est à noter que

incomplète et la méthode de pondération fondée sur la stratification synthétique double. Étant donné qu'on utilise deux ensemble différents de variables communes pour pondérer les grands sous-échantillons ainsi que pour effectuer l'appariement statistique, nous obtenons quatre ensemble de poids de calage pour chaque cycle de simulation par rapport au sous-échantillon de petite taille, qui donne à son tour, dans le cas de chaque cycle de simulation, quatre tableaux estimés à double entrée différents composés de X et Z . Afin de faciliter les calculs, nous avons utilisé la mesure de distance quadratique dans l'estimation de calage, ce qui implique que chaque cellule estimée correspond à une estimation de régression générale. Enfin, nous avons pris les moyennes et les variances de ces tableaux à double entrée au cours des 500 simulations. Les résultats sont indiqués dans les tableaux 5 à 8.

Tableau 5

Stratification double incomplète combinée au premier ensemble de variables communes					
oui					
447 ⁽⁹⁶⁾	232 ⁽⁹⁷⁾	89 ⁽²⁸⁾	25 ⁽²¹⁾	59 ⁽⁴⁹⁾	852 ⁽¹⁷⁾
non					
61 ⁽⁷⁹⁾	104 ⁽⁹⁰⁾	11 ⁽²¹⁾	11 ⁽¹⁹⁾	46 ⁽⁴⁶⁾	233 ⁽¹⁷⁾
total					
508 ⁽²³⁾	336 ⁽¹⁹⁾	100 ⁽⁸⁾	36 ⁽³⁾	105 ⁽¹⁰⁾	1 085
Stratification double incomplète combinée au deuxième ensemble de variables communes					
oui					
447 ⁽⁷⁵⁾	231 ⁽⁷⁴⁾	89 ⁽¹⁷⁾	25 ⁽²⁰⁾	59 ⁽⁴²⁾	851 ⁽¹⁷⁾
non					
61 ⁽⁵⁸⁾	105 ⁽⁶⁵⁾	11 ⁽¹²⁾	11 ⁽¹⁹⁾	46 ⁽³⁸⁾	234 ⁽¹⁷⁾
total					
508 ⁽²³⁾	336 ⁽¹⁹⁾	100 ⁽⁸⁾	36 ⁽³⁾	105 ⁽¹⁰⁾	1 085

Tableau 7

Stratification double synthétique combinée au premier ensemble de variables communes					
oui					
447 ⁽⁷⁵⁾	231 ⁽⁷⁴⁾	89 ⁽¹⁷⁾	25 ⁽²⁰⁾	59 ⁽⁴²⁾	851 ⁽¹⁷⁾
non					
61 ⁽⁵⁸⁾	105 ⁽⁶⁵⁾	11 ⁽¹²⁾	11 ⁽¹⁹⁾	46 ⁽³⁸⁾	234 ⁽¹⁷⁾
total					
508 ⁽²³⁾	336 ⁽¹⁹⁾	100 ⁽⁸⁾	36 ⁽³⁾	105 ⁽¹⁰⁾	1 085
Stratification synthétique double combinée au deuxième ensemble de variables communes					
oui					
447 ⁽⁷⁰⁾	231 ⁽⁷⁰⁾	89 ⁽¹⁶⁾	25 ⁽¹⁸⁾	59 ⁽⁴⁰⁾	851 ⁽¹⁷⁾
non					
61 ⁽⁵²⁾	105 ⁽⁶⁰⁾	11 ⁽¹¹⁾	11 ⁽¹⁶⁾	46 ⁽³⁷⁾	234 ⁽¹⁷⁾
total					
508 ⁽²³⁾	336 ⁽¹⁹⁾	100 ⁽⁸⁾	36 ⁽³⁾	105 ⁽⁹⁾	1 085

Tableau 8

Stratification synthétique double combinée au deuxième ensemble de variables communes					
oui					
447 ⁽⁷⁰⁾	231 ⁽⁷⁰⁾	89 ⁽¹⁶⁾	25 ⁽¹⁸⁾	59 ⁽⁴⁰⁾	851 ⁽¹⁷⁾
non					
61 ⁽⁵²⁾	105 ⁽⁶⁰⁾	11 ⁽¹¹⁾	11 ⁽¹⁶⁾	46 ⁽³⁷⁾	234 ⁽¹⁷⁾
total					
508 ⁽²³⁾	336 ⁽¹⁹⁾	100 ⁽⁸⁾	36 ⁽³⁾	105 ⁽⁹⁾	1 085

L'ensemble de données comprend 1 085 dossiers dont on a observé les variables suivantes: âge (six catégories: 15-24, 25-34, 35-44, 45-54, 55-64, 65 et plus), sexe (deux catégories: homme, femme), possession d'une maison (deux catégories: oui, non), occupation (cinq catégories: emploi, entretien ménager, études, travail bénévolé, autre) et santé (deux catégories: oui, non). Aux fins de l'étude de simulation, cet ensemble de données est considéré comme une population finie. On présuppose que les totaux de population relatifs à l'âge et au sexe sont connus.

Afin de simuler les méthodes de pondération, nous avons exécuté un algorithme de Monte Carlo. Nous avons notamment produit 500 échantillons, de manière indépendante les uns des autres, d'après un plan d'échantillonnage à deux phases. Dans la première phase, on recueille un échantillon aléatoire simple d'une taille de 20 500 personnes, avec remplacement. Dans cet échantillon, on observe l'âge, le sexe et la possession d'une maison. Dans la deuxième phase, l'échantillon recueilli lors de la première phase est divisé de façon aléatoire en deux grands sous-échantillons d'une taille de 10 000 personnes chacun et un petit sous-échantillon de 500 personnes; dans l'un des grands échantillon, on observe l'occupation (représentée par X), dans l'autre, on observe la santé (représentée par Z), tandis que dans le petit échantillon, on observe à la fois l'occupation et la santé. Lors de chaque observation, nous avons estimé le tableau à double entrée composé de X et Z d'après quatre méthodes de pondération que nous allons présenter ci-après.

L'échantillon de la première phase est pondéré au moyen d'un croisement du sexe et de l'âge qui est utilisé comme variable de contrôle. Il s'agit là uniquement d'une poststratification avec douze poststrates. D'après ces poids, les totaux de population peuvent être estimés dans le cas de toutes les variables observées dans l'échantillon de la première phase et des croisements effectués entre ces variables. En particulier, nous pouvons reproduire les totaux de population dans le cas de combinaisons de l'âge et du sexe, et obtenir des totaux de population estimés dans le cas de croisements de variables relatives à l'âge, au sexe et à la possession d'une maison. Maintenant, nous distinguons deux ensembles de variables communes pour pondérer les grands sous-échantillons, ainsi que pour obtenir une estimation pour le tableau synthétique à double entrée composé de X et Z . Le premier ensemble est une combinaison de l'âge et du sexe (12 catégories) et le deuxième une combinaison de l'âge, du sexe et de la possession d'une maison (24 catégories). Dans le cas de chaque simulation, cela donne deux estimations différentes pour les dénombrements marginaux, c'est-à-dire deux estimations différentes pour les totaux de population de X et Z (il est à noter que les deux estimations sont fondées sur une poststratification) et deux estimations différentes pour le tableau synthétique à double entrée. Afin de pondérer le petit sous-échantillon, nous faisons la distinction entre la méthode de pondération fondée sur la stratification double

3.2 Stratification double synthétique

Dans la présente section, nous considérons un autre estimateur pour le tableau YZ qui a fait appel également aux (grands) échantillons comme sources d'informations auxiliaires. Cependant, plutôt que d'utiliser des dénombremens estimés marginaux comme informations auxiliaires, on utilise des dénombremens synthétiques estimés. Soit B le coefficient de régression de population relatif à Y et C , qui est estimé d'après le premier (grand) échantillon :

$$B = \left(\sum_{k=1}^{ken_1} v_{1k} c_k c_k' \right)^{-1} \left(\sum_{k=1}^{ken_1} v_{1k} c_k y_k' \right).$$

De façon analogue, soit A le coefficient de régression de population relatif à Z et C , qui est estimé d'après le deuxième (grand) échantillon :

$$\hat{A} = \left(\sum_{k=1}^{ken_2} v_{2k} c_k c_k' \right)^{-1} \left(\sum_{k=1}^{ken_2} v_{2k} c_k z_k' \right).$$

Il est à noter que ces coefficients de régression estimés sont fondés sur les poids de calage de deuxième phase, plutôt que sur les poids d'inclusion. S'il y a une constante a qui fait que $a' c_k = 1$ dans le cas de toutes les personnes k , nous avons alors toujours $1' B' c_k = 1' A' c_k = 1$ dans le cas de toutes les personnes k . Maintenant, inspirés par la décomposition donnée par (3), c'est-à-dire,

$$\sum_{k=1}^N y_k z_k' = B' \sum_{k=1}^N (c_k c_k') A +$$

$$\sum_{k=1}^N (y_k - B' c_k)(z_k - A' c_k)',$$

nous proposons l'estimation du tableau à double entrée en deux étapes. Dans la première étape, le premier terme du côté droit est estimé en remplaçant les coefficients de régression de population B et A par leurs estimations B' et A' . En outre, nous proposons d'estimer $\sum_{k=1}^N c_k c_k'$ d'après l'estimation combinée

$$\sum_{k=1}^N c_k c_k' = \gamma \sum_{k=1}^{ken_1} v_{1k} (c_k c_k') + (1 - \gamma) \sum_{k=1}^{ken_2} v_{2k} (c_k c_k'),$$

où v_{1k} et v_{2k} représentent les poids (de deuxième phase) des premier et deuxième échantillons, et $\gamma \in [0, 1]$. Finalement, le premier terme est estimé d'après $B' \sum_{k=1}^N \hat{A}$. Jusqu'à maintenant, nous n'avons pas utilisé le petit échantillon (troisième). Si on veut, les estimations de B , A et $\sum_{k=1}^N c_k c_k'$ peuvent être légèrement améliorées en utilisant également l'échantillon de petite taille.

Dans la deuxième étape, on estime le tableau complet à double entrée composé de Y et Z en pondérant le petit échantillon d'après l'estimateur de calage

3.3 Une étude de simulation : intégration d'enquêtes-ménages

Il s'ensuit que les dénombremens marginaux de cellule du tableau estimé à double entrée sont les estimateurs à deux phases pour les totaux de population de Y et Z tels que nous les avons définis dans la section 3.1.

$$\sum_{k=1}^{ken_3} w_{3k} y_k = B' \left(\gamma \sum_{k=1}^{ken_1} v_{1k} c_k + (1 - \gamma) \sum_{k=1}^{ken_2} v_{2k} c_k \right) = B' \left(\sum_{k=1}^{ken_1} v_{1k} c_k \right) = B' \left(\sum_{k=1}^{ken_1} v_{1k} c_k c_k' \right) a = t_y'.$$

De la même façon, en postmultipliant les deux côtés par $1'$, nous obtenons un estimateur pour le total de population de

$$\sum_{k=1}^{ken_3} w_{3k} z_k' = \left(\gamma \sum_{k=1}^{ken_1} v_{1k} c_k' + (1 - \gamma) \sum_{k=1}^{ken_2} v_{2k} c_k' \right) \hat{A} = a' \left(\sum_{k=1}^{ken_2} v_{2k} c_k c_k' \right) \hat{A} = t_z'.$$

Le premier terme du côté droit est une estimation pour le tableau synthétique à double entrée. Cette estimation est presque non biaisée dans le cas du tableau YZ si l'hypothèse d'indépendance conditionnelle est confirmée. Nous signalons que ce type d'estimateur est obtenu essentiellement en appliquant la méthode d'appariement statistique avec contraintes (voir, par ex., Barr et Turner 1980, Rodgers 1984 ou Rubin 1986). Le deuxième terme est un terme de correction utilisé pour obtenir une estimation presque non biaisée pour le tableau YZ sans l'hypothèse précitée. S'il existe une constante a qui fait que $a' c_k = 1$ dans le cas de tous les éléments échantillonnés, nous obtenons alors, en prémultipliant les deux côtés de (8) par $1'$, l'estimateur suivant pour le total de population de Z :

$$\sum_{k=1}^{n_3} w_{3k} (y_k - B' c_k)(z_k - A' c_k)' + \sum_{k=1}^{n_3} w_{3k} z_k' = B' \sum_{k=1}^N c_k c_k' \hat{A} +$$

assujéti au troisième ensemble de contraintes (voir section 2.2), où B , A et $\sum_{k=1}^N c_k c_k'$ sont remplacés par leurs estimations B , A et $\sum_{k=1}^N c_k c_k'$. L'estimateur qui est obtenu correspond à

Dans la présente sous-section, nous allons comparer les méthodes de pondération que sont la stratification double incomplète (présentée en 3.1) et la stratification double synthétique (présentée en 3.2), à l'aide d'une étude de simulation. Nous utilisons un ensemble de données qui provient d'une étude pilote de la Dutch Household Survey on Living Conditions; voir van Tuinen (1995).

échantillons respectivement. Ces totaux de population sont estimés en deux phases. Dans la première phase, les deux grands échantillons sont pondérés en utilisant X comme un ensemble de variables de contrôle. Cela implique que les deux grands échantillons soient pondérés de manière à ce qu'ils reproduisent les totaux de population connus de X , qui sont représentés par t_x . D'après ces poids, une estimation regroupée des totaux de population de U donne

$$\hat{t}_u = \lambda \sum_{k \in n_1} w_{1k} u_k + (1 - \lambda) \sum_{k \in n_2} w_{2k} u_k,$$

où w_{1k} et w_{2k} représentent les poids de calage (de première phase) des premier et deuxième échantillons, et $\lambda \in [0, 1]$. Dans la deuxième phase, les deux échantillons sont pondérés à nouveau en utilisant simultanément X et U comme variables de contrôle. Soient v_{1k} et v_{2k} les poids de calage de cette deuxième phase. Les estimateurs pour les totaux de population de X et Z qui sont obtenus peuvent être considérés comme des estimateurs de calage en deux phases (voir Renssen et Nieuwenbroek 1997, section 6). Ces estimateurs sont représentés respectivement par \hat{t}_y et \hat{t}_z :

$$\hat{t}_y = \sum_{k \in n_1} v_{1k} y_k \text{ et } \hat{t}_z = \sum_{k \in n_2} v_{2k} z_k.$$

Nous remarquons que les deux estimateurs sont fondés sur un ensemble similaire de variables de contrôle. Si l'ensemble commun de variables est volumineux, on peut considérer l'utilisation d'un sous-ensemble plus petit pour pondérer les deux échantillons. En général, le sous-ensemble servant à pondérer le premier échantillon peut différer du sous-ensemble utilisé pour pondérer le deuxième échantillon. Cependant, nous présupposons dorénavant que les deux (grands) échantillons sont pondérés d'après le même ensemble de variables de contrôle.

Le tableau à double entrée composé de Y et Z peut être estimé par la pondération du troisième échantillon (petit échantillon), en utilisant simultanément Y et Z comme variables de contrôle, c'est-à-dire

$$\hat{t}_{yz} = \sum_{k \in n_3} w_{3k} (y_k z_k),$$

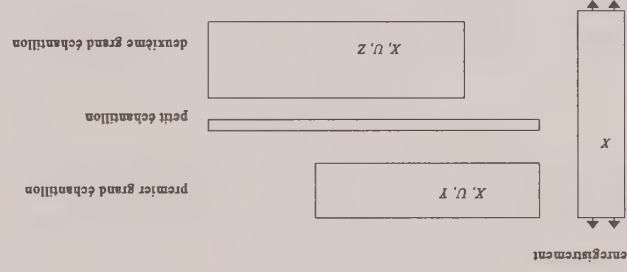
où les poids de calage w_{3k} satisfont aux contraintes

$$\sum_{k \in n_3} w_{3k} y_k = \hat{t}_y \text{ et } \sum_{k \in n_3} w_{3k} z_k = \hat{t}_z.$$

Il s'agit d'une stratification incomplète double, où les totaux de population inconnus de Y et Z sont remplacés par leurs estimations. Ces ensembles de contraintes permettent d'obtenir des estimations justes des dénombrements marginaux du tableau YZ , si les variables communes C présentent une forte corrélation avec Y et Z .

l'entremise de variables communes et en utilisant des informations auxiliaires tirées d'un échantillon de petite taille. Dans la présente section, nous adaptons cette méthode en combinant deux échantillons indépendants. Nous considérons un enregistrement complet de personnes, deux enquêtes sur échantillon à grande échelle et une enquête sur échantillon à petite échelle. L'enregistrement contient un ensemble limité de variables comme le sexe, l'âge, la région et l'état civil. Ces variables sont désignées par X . Dans l'un des grands échantillons, on observe les variables Y , U et X , tandis que dans l'autre grand échantillon, on observe les variables Z , U et X . Dans l'échantillon de petite taille, on observe toutes les variables, l'échantillon de petite taille peut provenir d'une enquête à petite échelle effectuée spécialement aux fins de la combinaison, ou du chevauchement des deux grandes enquêtes. Dans la figure 1, les sources de données sont indiquées schématiquement. Pour plus de commodité, on présume que tous les échantillons correspondent à des unités différentes, c'est-à-dire que l'on présume qu'il n'y a pas de chevauchement des échantillons.

Figure 1. Aperçu des diverses sources de données



Les variables X et U sont réparties dans $C = (X, U)$, où X désigne l'ensemble de variables communes comprenant des totaux de population connus et U l'ensemble de variables communes comprenant des totaux de population inconnus. Tous les échantillons peuvent être recueillis selon un plan d'échantillonnage complexe. On présume que Y et Z sont toutes deux des variables nominales; cependant, comme dans la section 2.5, les méthodes de pondération qui sont proposées peuvent aussi être appliquées à des variables Y et Z continues. Le but est d'estimer le tableau à double entrée composé par Y et Z . Nous considérons deux estimateurs. L'un est fondé sur une stratification incomplète double (analogue au premier ensemble de contraintes décrit dans la section 2.2), et l'autre est fondé sur un mélange entre l'appariement statistique et le calage (analogue au troisième ensemble de contraintes mentionné dans la section 2.2).

3.1 Stratification double incomplète

Tout d'abord, on estime les totaux de population relatifs à Y et Z au moyen des premier et deuxième (grands)

deux étapes. Lors de la première étape, les prédictions données par (7) sont calculées pour chaque (x_k, y_k) dans le premier enregistrement. Nous avons montré que les croisements entre les valeurs de X et ces valeurs prédites de Z peuvent être considérés comme des estimateurs à pondération. Cependant, les prédictions calculées n'ont en général pas de valeur réaliste; c'est pourquoi la première étape est suivie d'une deuxième. Dans cette deuxième étape, chaque valeur prédite de Z qui se trouve dans le premier enregistrement est remplacée par une valeur réelle de Z , tirée du deuxième enregistrement, qui se trouve le plus près dans (X, Z) d'après une certaine distance euclidienne.

2.5 Estimation de produits croisés dans le cas de variables X et Z continues

La caractéristique de cohérence du troisième ensemble de contraintes (section 2.2) est vérifiée également par rapport à des variables X et Z , pourvu qu'il y ait des constantes a , et a_z d'un ordre approprié, telles que $a_y y_k = 1$ et $a_z z_k = 1$ dans le cas de toutes les personnes k . Afin de voir cela, nous élargissons légèrement les résultats de la section 2.1. Notons d'abord que

$$a_y' B' x_k = a_y' \left(\sum_{i=1}^N y_i x_i' \right) \left(\sum_{i=1}^N x_i x_i' \right)^{-1} x_k =$$

$$a' \left(\sum_{i=1}^N x_i x_i' \right) \left(\sum_{i=1}^N x_i x_i' \right)^{-1} x_k = a' x_k = 1$$

(on présume toujours qu'il existe une constante a qui fait que $a' x_k = 1$ dans le cas de toutes les personnes k). De manière analogue, $a_z' A' x_k = 1$. Les équations équivalentes de (4) et (5) pour les variables continues sont obtenues facilement. Par conséquent, en prémultipliant les deux côtés de (III) avec a_y' , on obtient $\sum_{i=1}^N w_i z_i' = \sum_{i=1}^N w_i z_i'$, et en postmultipliant les deux côtés de (III) avec a_z' , on obtient $\sum_{i=1}^N w_i y_i' = \sum_{i=1}^N w_i y_i'$. Ainsi, le troisième ensemble de contraintes permet d'atteindre l'objectif de cohérence, c'est-à-dire l'équation de calage du premier ensemble de contraintes, dans le cas de variables X et Z assez générales. Voici deux exemples.

Dans le premier exemple, nous prenons $y_k = (1, y_{2k})'$ et $z_k = (1, z_{2k})'$, où l'on présuppose que y_{2k} et z_{2k} sont des variables continues. En prenant $a_y' = a_z' = (1, 0)'$, nous voyons que $a_y' y = a_z' z = 1$ dans le cas de toutes les personnes k . Le produit croisé de X et Z est égal à

$$\sum_{i=1}^N y_i z_i' = \begin{pmatrix} \sum_{i=1}^N y_{2i} & \sum_{i=1}^N y_{2i} z_{2i} \\ N & \sum_{i=1}^N z_{2i} \end{pmatrix},$$

du troisième ensemble de contraintes. Une élaboration de cet ensemble donne les quatre contraintes suivantes dans le cas de l'exemple qui nous intéresse:

$$\sum_{k=1}^N w_k = N, \quad \sum_{k=1}^N w_k y_{2k} = \sum_{k=1}^N y_{2k}, \quad \sum_{k=1}^N w_k z_{2k} = \sum_{k=1}^N z_{2k},$$

et

$$\sum_{k=1}^N w_k (y_{2k} z_{2k} - (y_{2k} - B_2' x_k)(z_{2k} - A_2' x_k)) =$$

où les coefficients de régression sont donnés par

$$B_2 = \left(\sum_{i=1}^N x_i x_i' \right)^{-1} \sum_{i=1}^N x_i y_{2i}$$

et par

$$A_2 = \left(\sum_{i=1}^N x_i x_i' \right)^{-1} \sum_{i=1}^N x_i z_{2i}.$$

Si l'on s'intéresse en particulier au coefficient de corrélation entre y_{2k} et z_{2k} , les contraintes suivantes pourraient également être prises en considération:

$$\sum_{k=1}^N w_k y_{2k} = \sum_{k=1}^N y_{2k} \quad \text{et} \quad \sum_{k=1}^N w_k z_{2k} = \sum_{k=1}^N z_{2k}.$$

Dans le deuxième exemple, nous supposons que $y_k = (1, y_{2k})'$, où y_{2k} peut être continue et où z_k est normale avec un nombre de catégories q . En prenant $a_y' = (1, 0)'$ et $a_z' = l$, où l est un vecteur de un ordre approprié, nous constatons que $a_y' y_k = a_z' z_k = 1$ dans le cas de toutes les personnes k . Le produit croisé de X et Z est

$$\sum_{i=1}^N y_i z_i' = \begin{pmatrix} \sum_{i=1}^N y_{2i} & \sum_{i=1}^N y_{2i} z_{2i} \\ N & \sum_{i=1}^N z_{2i} \end{pmatrix},$$

où C_h représente l'ensemble d'éléments de population appartenant à la catégorie h de Z , et N_h la taille de C_h . Nous nous assurons que les poids de calage relatifs au troisième ensemble de contraintes satisfont aux équations «marginales» de calage. $\sum_{k=1}^N w_k z_k' = \sum_{k=1}^N z_k' = (N_1 \dots N_q)'$ et $\sum_{k=1}^N w_k y_{2k} = \sum_{k=1}^N y_{2k}$, qui peuvent être toutes deux intéressantes en ce qui a trait aux exigences de cohérence.

3. COMBINAISON D'ÉCHANTILLONS INDÉPENDANTS PAR L'ENTREMISE DE VARIABLES COMMUNES

Dans la section précédente, nous avons présenté une méthode permettant de combiner deux enregistrements par

enregistrés. Les coefficients de régression partielle devraient être estimés à partir de la troisième source. Nous proposons

$$\hat{a}_1 = A - B\hat{a}_2$$

et

$$\hat{a}_2 = \left[\sum_{k=1}^N (Y_k - B'x_k)(Y_k - B'x_k)' \right]^{-1} \times \left[\sum_{k=1}^N w_k (Y_k - B'x_k)(z_k - A'x_k)' \right]$$

où w_k représente des poids de calage qui sont décrits dans la section 2.2. D'après ces estimations, nous définissons de nouvelles prédictions pour les valeurs Z:

$$\hat{z}_k = \hat{a}_1'x_k + \hat{a}_2'y_k = A'x_k + \hat{a}_2'(Y_k - B'x_k), k = 1, \dots, N. (7)$$

Ces nouvelles prédictions sont égales aux anciennes (voir section 2.1), un terme de correction en plus. Ce terme de correction dépend de la différence entre la valeur de son (ancienne) prédiction. Cependant, on peut considérer cela comme une tentative d'amélioration de la prédiction relative à Z et, plus important encore, comme un moyen de reconstruire l'estimateur à pondération d'après le troisième ensemble de contraintes (section 2.2). En effet, l'égalité suivante est vérifiée:

$$\sum_{k=1}^N Y_k \hat{z}_k' = \sum_{k=1}^N (B'x_k)(A'x_k)' + \sum_{k=1}^N w_k (Y_k - B'x_k)(z_k - A'x_k)'.$$

Il s'agit là uniquement de l'estimateur à pondération établi d'après le troisième ensemble de contraintes lorsque les poids de calage correspondants sont utilisés pour estimer α_2 . Il est facile de montrer que

$$\sum_{k=1}^N x_k \hat{z}_k' = \sum_{k=1}^N x_k z_k' = \sum_{k=1}^N x_k z_k'.$$

Ainsi, le tableau XZ peut également être reconstruit. Au début de la présente section, nous avons présupposé que le deuxième enregistrement est la source donnée. Ce choix est arbitraire. Si nous avions imputé les valeurs de Y au lieu des valeurs de Z, nous aurions obtenu une estimation identique pour le tableau YZ. En outre, le tableau XY aurait pu être reconstruit.

Les nouvelles prédictions relatives aux valeurs de Z peuvent être utilisées pour l'imputation. Singh et coll. (1993) donnent des algorithmes pour l'imputation en utilisant des modèles de régression. Ces valeurs de Z peuvent être imputées dans le premier enregistrement en

tableaux estimés à double entrée ou de rapports linéaires estimés plus généraux, mais également d'enregistrements complets dans lesquels sont inscrites simultanément les deux variables. Les utilisateurs de statistiques trouvent ce genre de bases de données complètes faciles à analyser. La création de tels enregistrements enrichis peut être considérée comme un cas spécial d'imputation. Un des enregistrements sert de source hôte ou de receveur, tandis que l'autre enregistrement sert de source donneuse. En admettant que le deuxième enregistrement est la source donneuse, la difficulté consiste à imputer dans le premier enregistrement les valeurs de Z tirées du deuxième enregistrement, en utilisant les informations contenues dans le tableau estimé à double entrée que nous avons décrit dans la section 2.2 comme informations auxiliaires. Des difficultés d'appariement statistique dans l'utilisation de données tirées d'une troisième source ont déjà été décrites par Rubin (1986) et Paass (1986). Singh et coll. (1993) présentent un examen de leurs méthodes. En outre, ils proposent certaines modifications des méthodes de Rubin (1986) et Paass (1986). Notre propre méthode d'imputation est fondée sur la méthode de régression proposée par Rubin (1986) et par Singh et coll. (1993).

où A est donné par (2), nous définissons de nouvelles prédictions de ces variables au moyen du modèle de régression élargi

$$\hat{z}_k = \alpha_1'x_k + \alpha_2'y_k, k = 1, \dots, N,$$

avec

$$\begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \left[\sum_{k=1}^N \begin{pmatrix} x_k x_k' & x_k y_k' \\ y_k x_k' & y_k y_k' \end{pmatrix} \right]^{-1} \left[\sum_{k=1}^N \begin{pmatrix} x_k z_k' \\ y_k z_k' \end{pmatrix} \right].$$

A l'aide de résultats bien connus concernant les coefficients de régression partielle du modèle linéaire général (voir, par ex., Seber 1977), α_1 et α_2 peuvent être exprimés sous la forme

$$\alpha_1 = A - B\alpha_2$$

et sous la forme

$$\alpha_2 = \left[\sum_{k=1}^N (Y_k - B'x_k)(Y_k - B'x_k)' \right]^{-1} \times$$

$$\left[\sum_{k=1}^N (Y_k - B'x_k)(z_k - A'x_k)' \right],$$

où B et A sont donnés par (1) et (2) respectivement. Ils peuvent être calculés à partir des premier et deuxième

À l'aide de cet enregistrement synthétique, de $r = 34\ 000$, on peut calculer le tableau synthétique à double entrée. Le résultat est présenté dans le tableau 1. Ce tableau peut être considéré comme une première approximation de la distribution statistique réelle entre l'âge et la possession d'une voiture. Une condition suffisante pour une approximation étroite est l'homogénéité en ce qui a trait à l'âge ou à la possession d'une automobile à l'intérieur de toutes les adresses, c'est-à-dire que toutes les personnes vivant à une même adresse devraient être soit dans la même catégorie d'âge, soit dans la même catégorie de possession d'une voiture. Dans le cas de la plupart des adresses à plus d'une personne, cette proposition semble peu probable. Il s'ensuit des équations (4) et (5) que les totaux de ligne et de colonne du tableau 1 correspondent aux dénombrements (marginaux) de population réels relatifs à l'âge et la possession d'une automobile respectivement.

À l'aide d'un échantillon aléatoire simple de $n = 1\ 000$ personnes, les dénombrements de population sont estimés au moyen d'un estimateur de régression générale. Trois ensembles de variables auxiliaires sont utilisés, conformément aux trois ensembles de contraintes mentionnés dans la section précédente. Les tableaux estimés sont indiqués ci-après (pour plus de commodité, nous avons utilisé la mesure de distance quadratique: $G(w_k/d_k) = (w_k/d_k - 1)^2$). Les écarts-types estimés correspondants sont indiqués entre parenthèses. Ces écarts-types estimés sont fondés sur les formules de variance habituelles de l'estimateur de régression générale (voir Særdal et coll. 1992, chap. 6).

Tableau 1
Taux de population synthétiques dans le cas de croisements de variables relatives à l'âge et à la possession d'une automobile

	1	2	3	4	5	6	total
oui	3 461	1 659	5 739	10 770	6 536	3 334	31 499
non	9 827	4 692	7 902	17 102	6 424	5 389	51 336
total	13 288	6 351	13 641	27 872	12 960	8 723	82 835

Dans le tableau 2, les dénombrements de population sont estimés d'après la stratification incomplète double ordinaire (Bethlehem et Keller 1987). Il n'y a pas de jeunes (catégories d'âge 1 et 2) qui possèdent une automobile dans l'échantillon, ce qui est une constatation susceptible d'être représentative de la population; ces cellules sont donc estimées à zéro. En raison des exigences de convergence, c'est-à-dire le premier ensemble de contraintes, les dénombrements estimés de jeunes ne possédant pas une automobile équivalent aux dénombrements de cellule marginaux correspondants. Une façon d'essayer d'améliorer le tableau 2 consiste à utiliser l'adresse à variable commune dans la méthode de pondération. Dans le tableau 3, les estimations relatives aux cellules sont indiquées d'après le deuxième ensemble de contraintes. Comme on l'a déjà

mentionné dans la section précédente, les totaux estimés de ligne et de colonne peuvent différer des dénombrements de population réels. Une comparaison entre le tableau 2 et le tableau 3 montre que ces différences peuvent être importantes. En outre, presque tous les dénombrements de cellule estimés plus faibles que les dénombrements de cellule estimés correspondants du tableau 3. Ainsi, le deuxième ensemble de contraintes donne des résultats assez insatisfaisants. Cela implique deux choses: 1) une cohérence des dénombrements de cellule marginaux estimés par rapport aux dénombrements de population connus correspondants; 2) des variances asymptotiques plus faibles de tous les dénombrements de cellule estimés. Les résultats sont indiqués dans le tableau 4. Les dénombrements marginaux estimés de cellule sont effectivement cohérents, et les écarts-types estimés sont, au plus, deux fois moins importants que les écarts-types estimés correspondants indiqués dans le tableau 2.

Tableau 2
Taux de population estimés dans le cas de croisements entre des variables relatives à l'âge et à la possession d'une automobile, d'après le premier ensemble de contraintes

	1	2	3	4	5	6	total
oui	0 ⁽⁰⁾	0 ⁽⁰⁾	4 968 ⁽⁴²³⁾	15 414 ⁽⁵⁴³⁾	7 518 ⁽⁴⁵⁸⁾	3 599 ⁽³⁷⁵⁾	31 499
non	13 288 ⁽⁰⁾	6 351 ⁽⁰⁾	8 673 ⁽⁴²³⁾	12 458 ⁽⁵⁴³⁾	5 422 ⁽⁴⁵⁸⁾	5 124 ⁽³⁷⁵⁾	51 336
total	13 288	6 351	13 641	27 872	12 960	8 723	82 835

Tableau 3
Taux de population estimés dans le cas de croisements entre des variables relatives à l'âge et à la possession d'une automobile, d'après le deuxième ensemble de contraintes

	1	2	3	4	5	6	total
oui	0 ⁽⁰⁾	0 ⁽⁰⁾	4 791 ⁽⁴³⁵⁾	13 826 ⁽⁸¹¹⁾	6 887 ⁽⁴⁹⁴⁾	3 421 ⁽³²¹⁾	28 923 ⁽¹⁰⁰⁵⁾
non	14 385 ⁽⁷⁸²⁾	7 012 ⁽⁵⁹⁵⁾	8 118 ⁽⁵⁶³⁾	12 893 ⁽⁷⁹⁶⁾	5 853 ⁽⁴⁶⁴⁾	5 654 ⁽³⁰⁶⁾	53 912 ⁽¹⁰⁰⁵⁾
total	14 385 ⁽⁷⁸²⁾	7 012 ⁽⁵⁹⁵⁾	12 908 ⁽⁶⁰³⁾	26 718 ⁽⁹⁵⁸⁾	12 739 ⁽⁴¹⁹⁾	9 074 ⁽¹⁷⁷⁾	82 835

Tableau 4
Taux de population estimés dans le cas de croisements entre des variables relatives à l'âge et à la possession d'une automobile, d'après le troisième ensemble de contraintes

	1	2	3	4	5	6	total
oui	0 ⁽⁰⁾	0 ⁽⁰⁾	5 501 ⁽³²⁶⁾	15 647 ⁽²²⁷⁾	6 898 ⁽¹⁷⁷⁾	3 453 ⁽⁷⁸⁾	31 499
non	13 288 ⁽⁰⁾	6 351 ⁽⁰⁾	8 139 ⁽³²⁶⁾	12 224 ⁽²²⁷⁾	6 062 ⁽¹⁷⁷⁾	5 270 ⁽⁷⁸⁾	51 336
total	13 288	6 351	13 641	27 872	12 960	8 723	82 835

2.4 Imputation de valeurs tirées d'un des enregistrements dans l'autre enregistrement

Jusqu'ici, nous avons élaboré une méthode de pondération pour estimer un tableau à double entrée réunissant deux variables qui sont inscrites dans deux enregistrements distincts. Souvent, on voudrait disposer non seulement de

Scheuren 1987). L'estimateur de calage défini d'après les deuxième et troisième ensembles de contraintes correspond à la poststratification, dans le sens que tous les croisements sont utilisés comme informations auxiliaires. À l'exception du cas où il existe un rapport linéaire parfait, soit entre X et X , soit entre Z et X , la méthode diffère de la poststratification complète par l'utilisation de totaux de population synthétiques plutôt que des dénombrements de résultats instables si certaines cellules d'échantillon contiennent seulement quelques observations. Dans de tels cas, la poststratification incomplète présente un intérêt sur le plan pratique. De façon analogue, l'estimateur de calage défini d'après les deuxième et troisième ensembles de contraintes peut être instable. Comme dans le cas de la poststratification incomplète, on peut considérer plutôt l'utilisation d'un croisement incomplet dans les contraintes.

2.3 Un exemple numérique

Nous illustrons l'utilisation de l'estimateur de calage dans le cas de trois ensembles de contraintes différents, à l'aide d'un exemple hypothétique. L'exemple est fondé sur des données réelles tirées d'un échantillon pour la Dutch National Travel Survey (1994). Le plan d'échantillonnage est grosso modo un échantillon en grappes d'adresses autopondérée. Toutes les personnes vivant à une adresse choisie sont incluses dans l'échantillon. La taille nette de l'échantillon est d'environ 80 000 personnes réparties dans 34 000 adresses. À partir de cet échantillon sont établis deux enregistrements hypothétiques d'environ $N = 80\ 000$ personnes. Dans l'un des enregistrements, on inscrit l'âge (dans six catégories) et dans l'autre, la possession d'une automobile (dans deux catégories). La variable commune entre les enregistrements est un nombre clé pour les adresses, ce qui donne $r = 34\ 000$ catégories dans le cas de la variable X . Dans le cas de cet exemple en particulier, le tableau synthétique à double entrée est simplifié comme suit:

$$\sum_{r=1}^N (B'x_k)'(A'x_k)' = \sum_{r=1}^N N_j \bar{y}_j \bar{z}_j',$$

où N_j représente la taille de l'adresse j , \bar{y}_j la moyenne des six catégories d'âge de l'adresse j , et \bar{z}_j la moyenne des deux catégories de possession d'une automobile de l'adresse j .

Afin de calculer le tableau synthétique à double entrée, les deux enregistrements sont combinés comme suit: premièrement, les enregistrements font l'objet d'un tri d'après le numéro clé pour les adresses; deuxièmement, on calcule les dénombrements des six catégories d'âge et des deux catégories de possession d'une automobile troisième-ment, chaque dénombrement d'adresses relatif à l'âge est relié à son dénombrement d'adresses correspondant relatif

En prémultipliant le troisième ensemble d'équations, des deux côtés, par l' , nous obtenons le premier ensemble de contraintes en ce qui a trait à Z , tandis qu'en postmultipliant le troisième ensemble des deux côtés par l , on obtient le premier ensemble de contraintes relativement à X . L'estimateur de calage qui en résulte peut être exprimé comme suit:

$$\hat{T} = \sum_{n=1}^N w_k (y_k z_k') = \sum_{n=1}^N (B'x_k)(A'x_k)' + \sum_{k=1}^K w_k (y_k - B'x_k)(z_k - A'x_k)'.$$

De toute évidence, cet estimateur suit la décomposition donnée par (3). Il est égal au tableau à deux entrées défini synthétiquement, un terme de correction en plus. Ce terme de correction est une estimation de calage de la différence relative à la distribution statistique réelle existant entre, d'une part, X et Z , et d'autre part le tableau à deux entrées défini synthétiquement. De façon analogue au deuxième ensemble de contraintes, dans le troisième ensemble, le nombre de contraintes non redondantes est borné par $p \times q$. Un cas spécial important est celui où $G(w_k/d_k) = (w_k/d_k - 1)^2$. Dans ce cas, chaque cellule estimée est une estimation de régression générale avec $(y_k z_k')$, $\text{vec}(B'x_k x_k' A)$, et $\text{vec}(y_k z_k' - (y_k - B'x_k)(z_k - A'x_k)')$ comme variables de contrôle, dans le cas des premier, deuxième et troisième ensembles de contraintes respectivement. Des formules analytiques pour la variance de plan d'échantillonnage de l'estimateur de régression générale sont données, par exemple, dans Särndal et coll. (1992, chap. 6). En fait, ces formules sont des approximations pour des tailles d'échantillon importantes. Dans Deville et Särndal (1992) sont décrites les conditions suffisantes dans lesquelles ces approximations sont valides, dans le cas d'estimateurs de calage en général.

Dans Deville et coll. (1993), la poststratification complète est décrite comme une méthode de calage pour laquelle on utilise tous les dénombrements de population relatifs à la classification croisée dans l'ensemble de contraintes. L'élaboration d'une poststratification complète donne l'estimateur de poststratification complète ordinaire, laquelle que soit la fonction de distance G . Solution de remplacement, la poststratification incomplète est décrite comme une méthode de calage dans laquelle on utilise une connaissance moins détaillée de tous les dénombrements de cellule dans l'ensemble de contraintes. L'estimateur de cellule défini d'après le premier ensemble de contraintes est un exemple couramment utilisé de poststratification incomplète. Plusieurs cas dans lesquels la poststratification incomplète est préférable à la poststratification complète ont été décrits. Deux de ces cas sont le manque d'informations relatives à la population et certains dénombrements de cellule zéro ou extrêmement petits (voir aussi Oh et

aux variables auxiliaires qui sont utilisées comme variables de contrôle dans l'estimateur. Une telle propriété de cohérence est intéressante si les informations auxiliaires sont utilisées tant pour la diffusion que pour la pondération. L'estimateur de calage est créé sous forme de généralisation de l'estimateur de régression générale (Deville et Särndal 1992, et Deville et coll. 1993).

Afin de donner un exemple précis, soit G une fonction fixe telle que définie dans Deville et coll. (1993) et considérons l'estimateur à pondération suivant dans le cas de notre tableau YZ :

$$(6) \quad \hat{T} = \sum_{n=1}^K w_k (y_k z_k'),$$

où w_k est un scalaire qui représente un poids attribué à la personne $k \in s$. Soit $d_k = \pi_k^{-1}$. Un estimateur de calage pour le tableau YZ utilise des poids qui sont obtenus en minimisant $\sum_{k=1}^K d_k G(w_k/d_k)$ par rapport à w_k , dans le cas de tout échantillon s . Nous considérons d'abord l'ensemble de contraintes suivant:

$$(I) \quad \sum_{n=1}^K w_k y_k = \sum_{N=1}^K y_k \quad \text{et} \quad \sum_{N=1}^K w_k z_k = \sum_{N=1}^K z_k.$$

Ce (premier) ensemble de contraintes utilise uniquement les dénombrements (marginiaux) relatifs à Y et Z . La variable commune X n'est pas utilisée. Une des $p + q$ équations est redondante, de sorte que pour résoudre le problème de la minimisation, une des équation peut être supprimée. Dans le cas de $G(w_k/d_k) = (w_k/d_k - 1)^2$, l'estimateur de calage qui est obtenu correspond à une stratification double incomplète, telle que définie dans Bethlehem et Keller (1987). En utilisant $G(w_k/d_k) = 1 + w_k/d_k (\log(w_k/d_k) - 1)$, on obtient l'estimateur classique itératif du quotient (voir, par ex., Oh et Scheuren 1987). Copeland, Petzmeier et Hoy (1987) ont comparé ces méthodes, d'après des données de la Current Population Survey. Ces auteurs en arrivent à la conclusion que les estimations obtenues à l'aide des deux méthodes sont très semblables. Dans Deville et coll. (1993), sont présentées deux autres fonctions de distance, qui sont particulièrement intéressantes compte tenu du problème lié aux poids extrêmes. L'estimation de tableaux à double entrée avec des contraintes sur les dénombrements marginaux est souvent effectuée dans le cas d'enquête sur échantillon. Souvent, les contraintes imposées aux dénombrements marginaux sont nécessaires pour deux raisons: premièrement, afin de réduire l'erreur et le biais d'échantillonnage, et deuxièmement, pour répondre à des exigences de cohérence relatives à des dénombrements de populations qui sont publiés.

Supposons que x_k est un score nominal qui comporte un nombre de catégories r . Étant donné qu'on dispose d'informations sur la population relatives aux croisements entre Y et X , et aux croisements entre Z et X , nous pourrions aussi considérer l'ensemble de contraintes suivant:

$$\sum_{n=1}^K w_k (y_k x_k') = \sum_{N=1}^K y_k x_k' \quad \text{et} \quad \sum_{n=1}^K w_k (z_k x_k') = \sum_{N=1}^K z_k x_k'.$$

Le nombre de contraintes non redondantes dans cet ensemble est égal à $r(p + q - 1)$. Dans le cas d'une valeur r plus grande, cet ensemble peut ne pas être applicable parce qu'il contient trop de contraintes par rapport à la taille de l'échantillon. L'ensemble pourrait présenter un intérêt pratique uniquement lorsque r est petit. Dans le reste du présent article, nous ne tiendrons pas compte de cet ensemble de contraintes.

En vue d'intégrer un ensemble important de variables communes à la méthode de pondération, nous considérons un ensemble de contraintes qui fait appel aux informations bidimensionnelles sur la population que nous avons dans le tableau synthétique:

$$(II) \quad \sum_{n=1}^K w_k (B' x_k) (A' x_k)' = \sum_{N=1}^K (B' x_k) (A' x_k)'.$$

Ce (deuxième) ensemble de contraintes est une application directe de la théorie des estimateurs de calage. Les totaux de population relatifs aux croisements entre $B' x_k$ et $A' x_k$ sont connus; ces combinaisons sont donc utilisées comme variables auxiliaires pour formuler l'ensemble de contraintes. Évidemment, dans le cas d'une valeur r importante, le nombre de contraintes non redondantes demeure limité par $p \times q$. Un inconvénient important des poids de calage qui sont obtenus est que ceux-ci ne reproduisent pas nécessairement les dénombrements (marginiaux) de population en ce qui a trait à Y et Z lorsqu'on applique ces poids à y_k et à z_k respectivement. Autrement dit, les poids de calage obtenus ne satisfont pas nécessairement au premier ensemble de contraintes. Il s'agit là d'un inconvénient important, surtout si cet ensemble est formulé en vue de respecter des exigences de cohérence.

Par conséquent, comme solution de remplacement, nous considérons un troisième ensemble de contraintes:

$$(III) \quad \sum_{n=1}^K w_k (y_k z_k' - (y_k - B' x_k)(z_k - A' x_k)) = \sum_{N=1}^K (B' x_k) (A' x_k)'.$$

En présupposant qu'il existe une constante a qui fait que $a' x_k = 1$ dans le cas de toutes les personnes k , cet ensemble de contraintes permet d'atteindre l'objectif de cohérence. Soit l un vecteur de un d'un ordre approprié, et rappelons-nous que $l' y_k = l' B' x_k = l' z_k = l' A' x_k = 1$ dans le cas de toutes les personnes k , que $B' \sum_{N=1}^K x_k = \sum_{N=1}^K 1 y_k$, et que

les scores x_k , $B'x_k$, et $A'x_k$. En fait, on peut ajouter soit y_k , soit z_k à l'enregistrement, mais dans le cas qui nous intéresse, cet ajout paraît superflu (voir prochain paragraphe). Lorsqu'il existe un vecteur a d'ordre r de nombres fixes qui fait que $a'x_k = 1$ dans le cas de toutes les personnes k , alors les totaux de population relatifs aux nouvelles variables $B'x_k$ et $A'x_k$ équivalaient aux totaux de population relatifs aux variables initiales correspondantes (voir, par ex., Bethlehém et Keller 1987). Cela peut être démontré facilement en prémultipliant d'abord les équations normales (1) et (2) par a' et en remplaçant par la suite $a'x_k = 1$ dans les équations qui sont obtenues. À partir de l'enregistrement synthétique, on peut définir un tableau synthétique à double entrée par $\sum_{k=1}^N (B'x_k)(A'x_k)'$. Ce tableau synthétique à double entrée peut être considéré comme une approximation de la distribution statistique (simultanée) $\sum_{k=1}^N y_k z_k'$. À l'aide des équations normales (1) et (2), on peut dériver les identités suivantes:

$$\sum_{k=1}^N (B'x_k)(A'x_k)' = \sum_{k=1}^N y_k (A'x_k)'$$

$$= \sum_{k=1}^N (B'x_k)z_k'$$

De toute évidence, les croisements entre $B'x_k$, $A'x_k$, y_k et z_k , ou entre $B'x_k$ et z_k donnent tous des tableaux synthétiques à double entrée identiques. Par conséquent, il suffit de considérer uniquement $\sum_{k=1}^N (B'x_k)(A'x_k)'$ et de supprimer y_k ou z_k dans l'enregistrement synthétique. La différence entre, d'une part, la distribution statistique réelle entre Y et Z et, d'autre part, son «approximation» synthétique peut être obtenue à l'aide de la décomposition suivante:

$$\sum_{k=1}^N y_k z_k' = \sum_{k=1}^N (B'x_k)(A'x_k)' +$$

$$\sum_{k=1}^N (y_k - B'x_k)(z_k - A'x_k)' \quad (3)$$

On remarquera la forte ressemblance avec la décomposition ordinaire de la variance effectuée dans l'analyse de régression (voir, par ex., Searle 1971). Lorsque $B'x_k = y_k$ ou $A'x_k = z_k$ dans le cas de toutes les personnes k , le tableau à double entrée dérivé de l'enregistrement synthétique équivaut à la distribution statistique simultanée réelle existant entre Y et Z .

Soit l un vecteur d'un ordre approprié consistant en des un, et $l'y_k = 1$ et $l'z_k = 1$ dans le cas de toutes les personnes k . S'il existe une constante a qui fait que $a'x_k = 1$ dans le cas de toutes les personnes k , nous avons alors également

dans le cas de toutes les personnes k , et, de façon analogue, $l'z_k = 1$, $A'x_k = 1$ dans le cas de toutes les personnes k . Il s'ensuit que

$$l' \sum_{k=1}^N (B'x_k)(A'x_k)' = \sum_{k=1}^N (A'x_k)' = \sum_{k=1}^N z_k' \quad (4)$$

$$\sum_{k=1}^N (B'x_k)(A'x_k)' l = \sum_{k=1}^N (B'x_k)' = \sum_{k=1}^N y_k' \quad (5)$$

Ainsi les totaux de ligne et de colonne du tableau synthétique à double entrée équivalent aux dénombrements marginaux de population correspondants en ce qui a trait à Y et Z .

Ce qu'il reste à considérer est la situation lorsque $a'x_k = 1$ dans le cas de toutes les valeurs k et d'une constante quelconque a . Cette condition est satisfaite lorsque X représente une variable nominale. De façon plus générale, cette condition est toujours satisfaite si le vecteur X peut être divisé en deux sous-vecteurs, dont l'un représente une variable nominale.

2.2 Formulation des contraintes dans l'estimation de calage

Supposons qu'un échantillon aléatoire s de taille n est tiré de la population finie $\Omega = \{1, \dots, N\}$ d'après un plan d'échantillonnage $p(s)$, de sorte que les probabilités d'inclusion de premier et de deuxième ordre $\Pr(k \in s) = \pi_k$ et $\Pr(k, l \in s) = \pi_{kl}$ sont strictement positives. Dans le cas de chaque $k \in s$, on observe le vecteur de scores (x_k, y_k, z_k) . On dispose de deux enregistrements distincts pour obtenir des informations auxiliaires. Le premier enregistrement contient, dans le cas de chaque $k \in \Omega$, des dossiers avec des scores relatifs à x_k et à y_k , tandis que le deuxième enregistrement contient, dans le cas de chaque $k \in \Omega$, des scores relatifs à x_k et à z_k . Le but est d'estimer le tableau YZ à partir de l'échantillon s , en utilisant des informations auxiliaires tirées des deux enregistrements. Il existe un large éventail d'estimateurs à pondération qui peuvent être utilisés en présence d'informations auxiliaires multivariées (1992). Dans Särndal, Swensson et Wretling (1992), l'estimateur de régression générale est décrit de manière approfondie. Il définit implicitement des poids d'échantillon qui reproduisent les totaux de population connus relatifs

Dans le présent article, nous procédons d'une manière différente.

2.1 Etablissement des totaux de population synthétiques

Soyent Y le niveau d'instruction avec un nombre de catégories p et Z l'emploi avec un nombre de catégories q . Nous avons alors y_k comme vecteur d'ordre p , qui représente des variables factices p . Chaque variable factice correspond à une catégorie précise et est égale à 1 si la personne k appartient à la catégorie en question, sinon elle est égale à 0. De façon analogue, z_k est un vecteur d'ordre q . En outre, X peut être le résultat d'un croisement (stratification) d'un certain nombre de caractéristiques (par ex. : sexe, âge, région, état civil, etc.). Les scores x_k sont un vecteur d'ordre r . Lorsque X consiste en une stratification complète, x_k représente un nombre de variables factices r . Dans le reste du présent article, r devrait être considéré comme un nombre important comparativement à $p \times q$. Les totaux de population dans le cas de X et Z représentent les distributions statistiques marginales relatives au niveau d'instruction et à l'emploi. À l'aide de la variable commune X , les prédictions pour Y et Z peuvent être définies à l'aide d'un modèle de régression linéaire multiple :

$$y_k = B'x_k, \quad k = 1, \dots, N,$$

et

$$z_k = A'x_k, \quad k = 1, \dots, N,$$

où B et A représentent les coefficients ordinaires de régression des moindres carrés qui satisfont aux équations

$$(1) \quad \begin{pmatrix} \sum_{k=1}^N x_k x_k' \\ \sum_{k=1}^N x_k y_k' \\ \sum_{k=1}^N y_k y_k' \end{pmatrix} B = \begin{pmatrix} \sum_{k=1}^N x_k x_k' \\ \sum_{k=1}^N x_k y_k' \\ \sum_{k=1}^N y_k y_k' \end{pmatrix} A$$

et

$$(2) \quad \begin{pmatrix} \sum_{k=1}^N x_k x_k' \\ \sum_{k=1}^N x_k z_k' \\ \sum_{k=1}^N z_k z_k' \end{pmatrix} A = \begin{pmatrix} \sum_{k=1}^N x_k x_k' \\ \sum_{k=1}^N x_k z_k' \\ \sum_{k=1}^N z_k z_k' \end{pmatrix} B$$

L'exposant « t » représente la transposition. Ce modèle est appelé modèle de probabilité linéaire (voir Maddala 1983, chap. 2). Il existe des modèles plus élégants, comme des modèles probit et logit, pour prédire des variables binaires. Cependant, nous ne nous intéressons pas aux prédictions elles-mêmes, mais plutôt aux totaux de population synthétiques de ces prédictions. Ces totaux semblent avoir des propriétés intéressantes lorsqu'on utilise le modèle de prédiction linéaire; c'est pourquoi il est possible de justifier l'utilisation de ce modèle. Il est à noter que B est calculé à partir du premier enregistrement et A à partir du deuxième. À l'aide de la variable commune X et des coefficients de régression B et A , nous créons un enregistrement synthétique qui contient un dossier pour chaque personne k , avec

Aux fins de cette présentation, il est pratique de décrire un estimateur de calage utilisé pour l'échantillon de petite taille, des informations auxiliaires étant obtenues de deux enregistrements distincts au lieu de deux grands échantillons distincts. Un des enregistrements contient des valeurs relatives à X et à Y , tandis que l'autre contient des valeurs relatives à X et à Z . Les sections 2.1 à 2.4 sont consacrées à des variables nominales Y et Z . Dans la section 2.1, les enregistrements sont utilisés pour obtenir une première estimation synthétique du tableau YZ par des méthodes d'imputation à régression. Nous montrons que ce tableau synthétique à double entrée possède des caractéristiques intéressantes. Dans la section 2.2, nous proposons un ensemble d'équations de calage qui servent à pondérer l'échantillon de petite taille d'après les caractéristiques en question. Nous décrivons brièvement le rapport qui existe entre cet ensemble et la poststratification complète, et entre cet ensemble et la poststratification incomplète. Nous donnons un exemple numérique dans la section 2.3. Le lien avec les méthodes d'appariement statistique, tel qu'il est présenté dans Singh et coll. (1993) est décrit dans la section 2.4. Le traitement des variables nominales Y et Z est excessivement restrictif. Dans la section 2.5, nous montrons que la méthode de pondération qui est proposée peut être appliquée également dans le cas de variables Y et Z continues ou dans le cas d'une variable Y continue et d'une variable Z nominale. Dans la section 3, nous modifions la méthode et utilisons des informations auxiliaires tirées de deux grands échantillons distincts plutôt que de deux enregistrements. À l'aide d'une étude de simulation, la méthode de pondération modifiée est comparée à la stratification double incomplète classique. Enfin, la section 4 est consacrée à quelques remarques conclusives.

2. COMBINAISON D'ENREGISTREMENTS PAR L'ENTREMISE DE VARIABLES COMMUNES

Prenons une population finie $\Omega = \{1, \dots, N\}$ composée d'un nombre N de personnes et supposons qu'il existe deux enregistrements concernant ces personnes. Le premier enregistrement contient, pour chaque personne k , un dossier contenant les scores y_k et x_k relatifs aux variables Y et X respectivement, tandis que le deuxième enregistrement contient, pour chaque personne k un dossier contenant les scores z_k et x_k relatifs aux variables Z et X respectivement. Comme on peut voir, la variable X est présente dans les deux enregistrements. Nous faisons remarquer que les dossiers des deux enregistrements correspondent à la même population finie. Le processus de fusion de ces deux enregistrements serait un appariement exact si l'on utilisait X pour comparer les dossiers contenus dans l'un des enregistrements avec ceux contenus dans l'autre, dans le but de déterminer quelles paires de dossiers ont trait à la même unité de population (voir Fellegi et Sunter 1969).

La méthode d'estimation qui est proposée ressemble étroitement à une méthode présentée par Singh et coll. (1993, section 2) et qui sert à estimer un coefficient de corrélation entre X et Z . Dans le présent article, nous présumons que ces variables sont unidimensionnelles. Cependant, notre méthode diffère de celle précitée par le fait qu'elle intègre à la méthode d'estimation le plan d'échantillonnage complexe de toutes les sources de données et qu'elle utilise les sources de données de grande taille d'une manière plus efficace dans l'estimation de paramètres de population provenant de la source de données de petite taille. Lorsque X et Z sont des variables nominales et qu'il n'y a pas de corrélation linéaire entre X et Z , notre méthode correspond à un tableau estimé à double entrée et Z qui correspond à un tableau estimé à double entrée qu'on obtiendrait à partir du fichier B si on imputait d'abord les valeurs de X . On obtient un résultat similaire lorsqu'il y a une corrélation parfaite entre Z et X .

Bien que la combinaison de sources de données distinctes par l'entremise de variables communes puisse être fructueuse d'un point de vue théorique, dans la pratique, on peut rencontrer des difficultés parce qu'on ne trouve pas facilement des variables communes au sens strict, principalement en raison de différences concernant les définitions, les méthodes d'observation et la période de référence. Ces difficultés peuvent être réduites si l'on harmonise les processus d'enquête à un stade initial. Une application prometteuse de l'utilisation de variables communes est le recours à des plans d'échantillonnage intégrés comme celui de la Dutch Household Survey on Living Conditions; voir van Tuinen (1995), Bakker et Winkels (1998), Winkels et Everaers (1998), et Hofmans (1998). Le questionnaire utilisé pour cette enquête comporte une structure à trois menus. Le premier menu contient des questions qui portent sur des aspects démographiques et socio-économiques, ainsi que sur le niveau d'instruction. Le deuxième menu contient des questions fondamentales auxquelles il est facile de répondre et qui portent sur tous les aspects pertinents des conditions de vie. Les questions contenues dans le troisième menu portent également sur les conditions de vie, mais elles sont plus exhaustives que les questions du deuxième menu. Afin de réduire le temps nécessaire pour répondre au questionnaire, le troisième menu est fractionné. Chaque répondant doit répondre à toutes les questions des premier et deuxième menus et remplir un sous-questionnaire du troisième menu. En raison du troisième menu, l'échantillon est divisé en sous-échantillons qui correspondent aux divers sous-questionnaires. Le plan d'échantillonnage de chaque sous-échantillon peut être décrit comme un échantillonnage à deux degrés pour l'estimateur de régression générale. Le présent article est structuré de la manière décrite ci-après. Le cadre théorique est présenté dans la section 2.

Dans l'ouvrage de Deville et Särndal (1992), l'estimation de calage est utilisée comme une méthode générale servant à pondérer des enquêtes sur échantillon, en tenant compte du plan complexe d'échantillonnage et d'informations auxiliaires obtenues de sources extérieures (voir aussi Deville, Särndal et Sautory 1993). L'utilisation d'informations auxiliaires, c'est-à-dire de variables de contrôle, vise principalement trois objectifs: réduire la variance d'échantillonnage, réduire le biais dû à la non-réponse et assurer la cohérence entre les estimations établies à partir de diverses sources par rapport aux variables de contrôle qui sont utilisées. Il existe une volumineuse documentation sur les méthodes de pondération utilisées dans le cadre d'enquêtes sur échantillon. Nous citons à cet égard Behllehem et Keller (1987), Alexander (1987), Lemaître et Dufour (1987), et Zieschang (1990).

Le présent article porte sur la façon d'estimer le produit croisé de X et Z (par ex., le tableau à double entrée de X et Z lorsque ces variables sont nominales, ou la covariance entre X et Z lorsque ces variables sont continues) à l'aide de méthodes d'appariement statistique et de l'estimation de calage. Nous présumons que deux fichiers de données A et B représentent deux vastes enquêtes sur échantillon, qui peuvent toutes deux être fondées sur un plan d'échantillonnage complexe. Afin de pondérer l'échantillon de l'enquête à petite échelle effectuée spécialement pour la combinaison de sources de données (fichier C), des informations auxiliaires sont tirées des deux grands échantillons. Il est peut-être difficile de déterminer si les grands échantillons devraient être considérés comme des sources d'informations auxiliaires pour l'échantillon de petite taille et vice-versa. L'appariement statistique permet d'intégrer un grand ensemble de variables X à la méthode d'estimation et de répondre à certaines exigences en matière de cohérence. La majeure partie de l'article est consacrée aux variables nominales X et Z en raison des caractéristiques de ces variables. Ainsi, nous montrons que les dénombrements marginaux du tableau YZ estimé correspondent toujours aux estimations des totaux de population de X et de Z lorsque l'estimateur de calage courant est appliqué en utilisant les variables X comme variables de contrôle dans le cas des premier et deuxième échantillons de grande taille respectivement. Cependant, la méthode qui est proposée est également applicable à des variables X et Z continues. Tout au long du présent article, nous présumons que X consiste en plusieurs variables, qui peuvent être nominales ou continues. Nous soutenons que lorsque les variables X présentent une forte corrélation avec X ou Z , notre méthode d'estimation donne des estimations relativement justes du produit croisé de X et Z , par exemple dans le cas du tableau complet YZ lorsque X et Z sont des variables nominales.

Utilisation de méthodes d'appariement statistique dans l'estimation de calage

ROBERT H. RENNSSEN¹

RÉSUMÉ

Le présent article porte sur un essai de mise en tableau croisé de deux variables nominales qui ont été recueillies de manière distincte à partir de deux échantillons indépendants de grande taille, et recueillies conjointement à partir d'un seul échantillon de petite taille. Dans le cadre de cet essai, on a présupposé que les échantillons de grande taille présentaient un grand ensemble de variables communes. La méthode d'estimation qui est proposée peut être considérée comme un mélange entre les méthodes de calage et l'appariement statistique. Grâce aux méthodes de calage, il est possible d'intégrer les plans d'échantillonnage complexes à la méthode d'estimation, afin de répondre à certaines exigences en matière de cohérence entre des estimations provenant de sources différentes, ainsi que pour obtenir des estimations plutôt non biaisées dans le cas du tableau à double entrée. Grâce aux méthodes d'appariement statistique, il est possible d'intégrer un ensemble relativement important de variables communes à l'estimation de calage, à l'aide de laquelle on peut améliorer la justesse du tableau à double entrée qui est estimé. La méthode d'estimation nous permet de mieux comprendre le biais qui accompagne généralement l'estimation du tableau à double entrée lorsque l'on utilise uniquement les échantillons de grande taille. Nous montrons l'utilité de la méthode d'estimation dans l'imputation des valeurs provenant d'un des grands échantillons (source donnée) à l'autre grand échantillon (source hôte). Bien que la méthode soit élaborée principalement pour les valeurs nominales X et Z , une modification mineure permet de l'appliquer également à des valeurs X et Z continues.

MOTS CLÉS: Cohérence entre les estimations; estimateur de régression générale; imputation; informations auxiliaires multidimensionnelles; tableau à double entrée.

1. INTRODUCTION

La plupart des enquêtes statistiques ont pour but d'obtenir de simples paramètres descriptifs relatifs à une population finie. Les estimations sont souvent présentées sous forme de tableaux comprenant des cellules qui contiennent les estimations de totaux de population ou de sous-groupes. Les données sont souvent recueillies d'après un vaste ensemble de variables, ce qui produit de nombreux résultats liés aux variables en question ainsi qu'aux rapports existant entre celles-ci. Afin d'économiser des ressources et de réduire le fardeau des répondants, les organismes qui s'occupent de statistique souhaitent réduire la taille des échantillons et la longueur des questionnaires. Ces organismes ont recours à des sources de données administratives et aux données d'enquêtes sur échantillon existantes qui sont menées à grande échelle, ou appliquent des plans d'échantillonnage à questionnaire fractionné (voir Raghunathan et Grizzle 1995). En raison de ces choix, les méthodes utilisées pour combiner des sources de données sont devenues des outils très prisés dans la production de données statistiques. La combinaison de sources de données peut être effectuée de nombreuses façons différentes; deux méthodes bien connues dans l'échantillonnage d'enquête sont l'appariement statistique et l'estimation de calage.

Singh, Mantel, Kinack et Rowe (1993) décrivent l'appariement statistique comme un cas particulier d'imputation qui comporte deux sources distinctes de micro-données

contenant différentes informations relatives à des unités différentes. Une des sources de données sert de fichier hôte ou de fichier receveur dans lequel sont imputées de nouvelles informations pour chaque enregistré à partir des données provenant de l'autre source, qui est le fichier donneur. Plus précisément, ces auteurs utilisent un fichier hôte A contenant des informations sur les variables (X , Y) et un fichier donneur B contenant des informations sur les variables (X , Z). La variable commune X peut être utilisée pour identifier des unités similaires dans les deux fichiers. En général, l'appariement statistique a trait à la difficulté que comporte l'achèvement des enregistrements dans le fichier A en imputant des valeurs pour Z à l'aide d'informations relatives au rapport existant entre X et Z qui proviennent du fichier B. Ces valeurs de Z imputées comportent un inconvénient important parce que le rapport réel existant entre X et Z peut être perdu en entier dans le fichier hôte qui est enrichi. Cet inconvénient correspond à ce qu'on appelle l'hypothèse d'indépendance conditionnelle entre Y et Z compte tenu de X . Afin de se débarrasser de cette hypothèse d'indépendance conditionnelle, Singh et coll. (1993) considèrent un troisième ensemble de données (fichier C) contenant des informations auxiliaires sur l'ensemble complet (X , Y , Z). Cet ensemble de données pourrait provenir, par exemple, d'une enquête spéciale menée à une petite échelle. Ces auteurs présentent plusieurs méthodes d'imputation permettant de compléter le fichier A par l'ajout de Z , à partir du fichier B, à l'aide de renseignements provenant de A, B et C et portant sur les

¹ Robert H. Rennsen, Department of Statistical Methods, Statistics Netherlands, P.O. Box 4481, 6401 CZ Heerlen, Netherlands.

$$\hat{d} = \begin{bmatrix} \hat{\beta}_{121} & 0,07851 \\ \hat{\beta}_{131} & 0,07558 \\ \hat{\beta}_{112} & -1,2918 \\ \hat{\beta}_{132} & -0,04813 \\ \hat{\beta}_{113} & -0,64214 \\ \hat{\beta}_{123} & 0,03884 \end{bmatrix}.$$

La matrice de covariance estimée, par notre analyse de régression non linéaire SAS, laquelle est basée sur le modèle initial qui présume que le taux d'erreur est fixe d'année en année et donc est le même pour toutes les années à l'étude, est définie ci-après.

Σ	β_{121}	β_{131}	β_{112}	β_{132}	β_{113}	β_{123}
β_{121}	0,000358	-4,7E-05	-3,5E-07	-2,6E-08	-3,9E-07	2,9E-07
β_{131}	-4,7E-05	0,000214	-1,7E-07	-5,2E-07	-1,4E-06	-2,8E-07
β_{112}	-3,5E-07	-1,7E-07	1,54E-06	2,14E-07	-2,3E-08	9,9E-10
β_{132}	-2,6E-08	-5,2E-07	2,14E-07	2,37E-06	-1,5E-08	-6,1E-08
β_{113}	-3,9E-07	-1,4E-06	-2,3E-08	-1,5E-08	2,4E-06	-8E-08
β_{123}	2,9E-07	-2,8E-07	9,9E-10	-6,1E-08	-8E-08	6,1E-06

La multiplication a priori et a posteriori du vecteur d par la matrice de covariance estimée donne une variance estimée pour AVE_{BLS} qui est de 6,72 E-6 pour 1989 et une erreur-type de 0,0026 (0,26 %), comme l'indique le tableau 4.

BIBLIOGRAPHIE

ABOWD, J., et ZELLNER, A. (1985). Estimated gross labor force flows. *Journal of Economic and Business Statistics*, 3, 253-283.

BAILLAR, B.A. (1968). Recent research in reinterview procedures. *Journal of the American Statistical Association*, 63, 41-63.

BIEMER, P.P., et FORSMAN, G. (1992). On the quality of reinterview data with application to the current population survey. *Journal of the American Statistical Association*, 87, 915-923.

BUREAU OF LABOR STATISTICS (1993). Overhauling the population survey. *Monthly Labor Review*, 116, 9. Washington DC: U.S. Government Printing Office.

BUREAU OF LABOR STATISTICS (1992). *Employment and Earnings*. 38, 8. Washington DC: U.S. Government Printing Office.

CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

FORMAN, G., et SCHREINER, I. (1991). The design and analysis of reinterview: an overview. Dans *Measurement Errors in Surveys*, (Eds. Paul Biemer et coll.). New York: John Wiley and Sons.

GASTWIRTH, J.L. (1987). The statistical precision of medical screening procedures: Application to polygraph and AIDs antibodies test data. *Statistical Science*, 2, 213-238.

HAUSMAN, J.A., et MORTON, S. (1994). Misclassification of a Dependent Variables in a Discrete Response Setting. Document de travail, Department of Economics, MIT, Cambridge.

HUI, S.L., et WALTER, S.D. (1980). Estimating the error rates of diagnostic tests. *Biometrics*, 36, 167-171.

HUGHES, J.J., et PERLMAN, R. (1984). *The Economics of Unemployment*. New York: Cambridge University Press.

LEVY, P., et LEMESHOW, S. (1980). *Sampling for Health Professionals*. California: Lifetime Learning Publications.

McKENNA, C.J. (1985). *Uncertainty and the Labor Market: Recent Developments in Job Search Theory*. New York: St. Martins Press.

POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.

POTERBA, J.M., et SUMMERS, L.H. (1995). Unemployment benefits and labor market transitions: a multinomial logit model with errors in classification. *The Review of Economics and Statistics*, 77, 207-216.

RAO, J.N.K., et THOMAS, D.R. (1991). Chi-squared tests with complex survey data subject to misclassification error. Dans *Measurement Errors in Surveys*, (Eds. Paul Biemer et coll.). New York: Wiley.

SCHREINER, I. (1980). Reinterview Results From the CPS Independent Reconciliation Experiment (second quarter 1978 through third quarter 1979). Note de service non publiée, U.S. Bureau of the Census, Mai 7, 1980.

SINCLAIR, M.D. (1994). Evaluating Reinterview Survey Methods for Measuring Response Errors. Thèse de doctorat, George Washington University, Septembre.

SINCLAIR M.D., et GASTWIRTH, J.L. (1996). On procedures for evaluating the effectiveness of reinterview survey methods: application to labor force data. *Journal of the American Statistical Association*, 91, 961-969.

SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flow estimate for classification error with application to data from the Canadian Labour Force Survey. *Journal of the American Statistical Association*, 90, 478-488.

U.S. BUREAU OF THE CENSUS (1963). *The Current Population Survey Reinterview Program: Some Notes and Discussion*. Document Technique no. 6. Washington, D.C.: U.S. Government Printing Office.

U.S. BUREAU OF THE CENSUS (1963). *Evaluating Censuses of Population and Housing*. Statistical Training Document #ISP-TR-5. Washington, D.C.: U.S. Government Printing Office, 1985.

VACEK, P.M. (1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, 41, 959-968.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

était négligeable. Même si ce n'est pas le cas, cette hypothèse simplifie grandement le calcul de la variance et saisit la majeure partie de la variation totale. Par ailleurs, cette hypothèse est corroborée par le fait que, compte tenu de la grande taille de l'échantillon annuel intégral qui est utilisé pour la CPS, la variance dans ces échantillons est négligeable lorsqu'on la compare à l'erreur d'échantillonnage associée aux estimations des taux d'erreur, lesquelles sont basées sur le petit échantillon de la réinterview non conciliée. En résumé, lorsque la substitution de U_R^y par U_{BLS}^y et de E_R^y par E_{BLS}^y dans l'expression (B.4) est complétée, nous présumons que U_{BLS}^y et E_{BLS}^y sont des valeurs fixes connues dans cette équation. Enfin, la variance d'échantillonnage, qui est associée à la différence entre la valeur corrigée et la valeur publiée et qui définit le biais dans l'estimation initiale, est calculée en faisant la somme des variances. Par conséquent, si nous présumons que la valeur publiée est exempte de variance d'échantillonnage, la variabilité d'échantillonnage associée à la différence ou au biais est tout simplement égale à la variabilité d'échantillonnage de la valeur corrigée.

ANNEXE TECHNIQUE C

Estimation des erreurs-types des taux de chômage corrigés

Pour une fonction complexe de plusieurs paramètres estimés, les estimations de la variance associée à cette fonction peuvent être calculées par approximation par série de Taylor, comme l'a proposé Wolter (1985). Supposons que le paramètre à étudier est $Y = G(\Theta)$, où Θ représente un n vecteur dimensionnel des paramètres de la population, $\Theta = \{\theta_1, \dots, \theta_n\}$. Si G possède des dérivées secondes continues qui se situent dans un intervalle acceptable pour

$$\hat{Y} - Y = A + R(\hat{\Theta}, \Theta)$$

où,

$$A = \sum_{k=1}^K \frac{\partial G(\Theta)}{\partial \theta_k} (\hat{\theta}_k - \theta_k) \quad R(\hat{\Theta}, \Theta) = \sum_{n=1}^K \sum_{i=1}^n (1/2!) \frac{\partial^2 G(\Lambda)}{\partial \theta_k \partial \theta_i} (\hat{\theta}_k - \theta_k)(\hat{\theta}_i - \theta_i) \quad \hat{\Theta} \leq \Lambda \leq \Theta. \quad (C.1)$$

Le terme qui reste est souvent considéré comme ayant peu de conséquence et est éliminé de la relation. Avec une approximation du premier ordre, Wolter présente,

$$MSE(\hat{Y}) = E[G(\hat{\Theta}) - G(\Theta)]^2 = \text{Var}(A)$$

$$= \sum_{n=1}^K \sum_{i=1}^n \frac{\partial G(\Theta)}{\partial \theta_k} \frac{\partial G(\Theta)}{\partial \theta_i} \text{Cov}(\hat{\theta}_k, \hat{\theta}_i)$$

$$= d \Sigma_{\hat{\Theta}} d^T \quad (C.2)$$

où d est un vecteur ligne de dimension n , formé des éléments

$$d_k = \left[\frac{\partial G(\Theta)}{\partial \theta_k} \right]. \quad (C.3)$$

Wolter désigne cet estimateur l'approximation du premier ordre de l'erreur quadratique moyenne (laquelle est égale à la somme de la variance d'échantillonnage et du biais de l'estimateur au carré). On peut obtenir des approximations d'ordre supérieur, en conservant des termes additionnels dans le développement. Pour l'estimation de la variance, nous remplaçons $\Sigma_{\hat{\Theta}}$ par la matrice de covariance estimée et nous évaluons d d'après les valeurs estimées de Θ . Dans le cas particulier qui nous intéresse, nous voulons estimer la variance associée à la fonction des estimations dans l'expression (C.4) indiquée ci-après.

$$G(\Theta) = G(\beta_{121}^y, \beta_{131}^y, \beta_{112}^y, \beta_{132}^y, \beta_{113}^y, \beta_{123}^y, U_{BLS}^y, E_{BLS}^y) =$$

$$\left\{ -U_{BLS}^y + E_{BLS}^y + \beta_{112}^y (U_{BLS}^y - \beta_{113}^y - E_{BLS}^y) \right. \\ \left. + \beta_{132}^y (U_{BLS}^y - \beta_{113}^y - E_{BLS}^y) + \beta_{123}^y (U_{BLS}^y - \beta_{112}^y - E_{BLS}^y) \right\} \\ \frac{+ \beta_{121}^y (U_{BLS}^y - \beta_{113}^y - E_{BLS}^y) + \beta_{113}^y (1 + \beta_{112}^y - \beta_{121}^y - \beta_{123}^y) + \beta_{112}^y (U_{BLS}^y - \beta_{113}^y - E_{BLS}^y)}{+ \beta_{121}^y (U_{BLS}^y - \beta_{113}^y - E_{BLS}^y) - \beta_{123}^y (U_{BLS}^y - \beta_{112}^y - E_{BLS}^y)} \quad (C.4)$$

Pour obtenir les estimations, nous présumons que les valeurs de U_{BLS}^y et E_{BLS}^y sont fixes (c'est-à-dire que la variance d'échantillonnage est négligeable). Si nous prenons les dérivées partielles de l'équation (C.4) pour les six taux d'erreur et que nous évaluons ces expressions en fonction des taux d'erreur estimés, nous obtenons un vecteur d qui dépend des estimations des taux d'erreur ainsi que des proportions de chômeurs et de personnes occupées publiées par le BLS pour chaque année à l'étude. Dans notre modèle initial qui présume que les taux d'erreur demeurent fixes d'une année à l'autre, ce vecteur d pour la période à l'étude ne varie, d'une année à l'autre, que pour les valeurs publiées. À titre d'exemple, pour l'année 1989, le vecteur d estimé à partir des taux de chômage et d'emploi publiés par le BLS (soit 0,0347 et 0,6329), est égal à:

$$E_1 = p_{2 \cdot 1} - p_{1 \cdot 1}, \quad E_2 = p_{21 \cdot} - p_{11 \cdot}.$$

À noter qu'il existe deux points distincts dans l'ensemble de solutions, pour une valeur positive ou négative de D ; cependant, une seule de ces valeurs donnera des estimations raisonnables. Les variances des estimateurs, calculées à partir de la matrice d'information asymptotique estimée, sont définies dans le document de Hui et Walter (1980).

ANNEXE TECHNIQUE B

Correction des taux de chômage publiés

Pour évaluer les répercussions des taux d'erreur estimés,

nous avons besoin d'une expression pour estimer les taux de prévalence réels (les quatre paramètres π), en fonction des taux d'erreur estimés et des taux de prévalence observés (ou fréquences dans l'échantillon), pour une enquête donnée. Nous présentons ici la formule utilisée pour ces calculs, laquelle peut également servir à calculer les taux corrigés du BLS en utilisant, comme valeurs observées, les taux de chômage et les taux d'emploi publiés par le BLS.

Cette expression est définie en (B.1).

À noter que, dans l'expression (B.1), nous avons supprimé l'indice g -ième des paramètres π , afin que l'expression représente les taux de prévalence parmi la population en général, hommes et femmes réunis. Pour cette étude, nous présumons que les taux d'erreur pour les hommes et les femmes sont égaux.

Dans l'analyse présentée ici, nous avons trois ensembles de valeurs observées, soit: deux taux de prévalence observés – établis à partir de l'échantillon de réinterview (lequel est un sous-échantillon de l'échantillon intégral de la CPS) et incluant les données non conciliées de la réinterview et les données conciliées de l'étude sur le biais dans les réponses – et les taux de prévalence publiés par le BLS à partir de l'enquête initiale intégrale (CPS). Nous nous concentrerons ici sur les premier et dernier ensembles de statistiques, à savoir les données non conciliées de la réinterview et les estimations publiées du BLS. Afin de maintenir une distinction entre ces deux ensembles, nous définissons,

$$(B.1) \quad \begin{bmatrix} \pi_{y1} \\ \pi_{y2} \end{bmatrix} = \begin{bmatrix} 1 - \beta_{121} - \beta_{131} - \beta_{113} & \beta_{112} - \beta_{113} \\ \beta_{121} - \beta_{123} & 1 - \beta_{112} - \beta_{132} - \beta_{123} \end{bmatrix}^{-1} \begin{bmatrix} \frac{n_{y..}}{n_{y1.}} - \beta_{113} \\ \frac{n_{y..}}{n_{y2.}} - \beta_{123} \end{bmatrix}.$$

$$(B.2) \quad U_R^y = \frac{n_{y1.}}{n_{y..}}, \quad E_R^y = \frac{n_{y2.}}{n_{y..}}.$$

comme étant les taux de prévalence observés de chômeurs et de personnes occupées, d'après les données non conciliées de la réinterview. Les taux de prévalence correspondants, publiés par le BLS et basés sur les données pondérées de l'enquête initiale intégrale (CPS), sont définis par U_{BLS}^y et E_{BLS}^y .

De même, le taux de chômage observé dans la population active, d'après l'échantillon de réinterview non conciliée, est représenté par UE_R^y , qui est égal à U_R^y divisé par $(U_R^y + E_R^y)$, alors que le taux de chômage publié par le BLS est représenté par UE_{BLS}^y .

Si on simplifie l'expression (B.1) en fonction des taux de prévalence observés lors de la réinterview, U_R^y et E_R^y , nous obtenons:

$$\begin{aligned} \pi_{y1} &= \left\{ U_R^y - \beta_{113} - \beta_{112} U_R^y + \beta_{113} \beta_{132} \right. \\ &\quad \left. - \beta_{123} U_R^y - \beta_{112} E_R^y + \beta_{113} \beta_{132} + \beta_{123} E_R^y \right\} \\ &\quad \left\{ 1 - \beta_{112} - \beta_{132} - \beta_{123} (1 + \beta_{132} + \beta_{113}) \right. \\ &\quad \left. - \beta_{131} (\beta_{112} + \beta_{132} - 1) + \beta_{123} \beta_{112} \right\}^{-1} \\ &\quad \left\{ 1 - \beta_{121} U_R^y + \beta_{121} \beta_{113} + \beta_{123} U_R^y - \beta_{123} E_R^y \right. \\ &\quad \left. + \beta_{122} \beta_{123} - \beta_{131} E_R^y + \beta_{131} \beta_{123} - \beta_{123} E_R^y \right\}^{-1} \\ &\quad \left\{ 1 - \beta_{112} - \beta_{132} - \beta_{123} (1 + \beta_{132} + \beta_{113}) \right. \\ &\quad \left. - \beta_{131} (\beta_{112} + \beta_{132} - 1) + \beta_{123} \beta_{112} \right\}^{-1}. \end{aligned} \quad (B.3)$$

À partir de l'expression (B.3), nous pouvons estimer le taux de chômage corrigé à partir de la réinterview, représenté par AE_R^y , égal à π_{y81} divisé par $(\pi_{y81} + \pi_{y82})$. À noter que AE_R^y peut être représenté par:

$$AE_R^y = \left\{ -U_R^y + E_R^y + \beta_{112} (U_R^y - \beta_{113} + E_R^y) \right. \\ \left. + \beta_{132} (U_R^y - \beta_{113}) + \beta_{123} (U_R^y - \beta_{113} - E_R^y) \right. \\ \left. + \beta_{113} (1 + \beta_{112} - \beta_{123}) + \beta_{112} (U_R^y + E_R^y - \beta_{113}) \right. \\ \left. + \beta_{121} (U_R^y + E_R^y - \beta_{123}) - E_R^y + \beta_{123} + \beta_{131} (U_R^y - \beta_{123}) \right\}^{-1} \quad (B.4)$$

Enfin, pour obtenir l'estimation corrigée du taux de chômage publié par le BLS, laquelle est représentée par AE_{BLS}^y , nous remplaçons U_R^y par U_{BLS}^y et E_R^y par E_{BLS}^y dans l'expression (B.4). À noter que les erreurs-types estimées des estimations pour AE_{BLS}^y qui ont été présentées à la section quatre, ont été calculées par une méthode d'approximation par série de Taylor (Wolter 1985). Nous avons d'abord présumé, pour ce faire, que la variance dans les estimations publiées de U_{BLS}^y et E_{BLS}^y

de comparaison exempte d'erreurs. La méthode prévoit l'utilisation de deux populations (ou sous-populations), auxquelles correspondent des taux de prévalence différents, pour estimer les paramètres. Les données d'une telle étude peuvent être résumées dans un tableau 2×2 comme celui illustré à la figure A qui suit. Ce tableau, lorsqu'il se rapporte à une sous-population précise, est indexé par la lettre g . Nous représentons par n_{ij}^g la fréquence des cas qui, dans la sous-population g , sont classés au premier test dans la catégorie i ($i = 1$ pour ceux qui possèdent la caractéristique à l'étude et $i = 2$ pour ceux qui ne l'ont pas) et, au deuxième test, dans la catégorie j ($j = 1$ ou 2). Supposons que π représente le taux de prévalence inconnu véritable de la caractéristique à l'étude et que α_r et β_r représentent les taux taux positif et négatif inconnus. Ces taux d'erreur sont indexés par la lettre r , où $r = 1$ correspond au résultat du premier test et $r = 2$, au résultat du deuxième (dans notre cas, $r = 1$ correspond à l'enquête initiale et $r = 2$ à la réinterview). Le taux taux positif, α_r , indique la probabilité que l'évaluation au r -ième test classe la personne comme étant positive alors qu'elle devrait en fait être classée parmi les négatifs. De même, le faux taux négatif, β_r , est la probabilité que l'évaluation au r -ième test classe le cas comme négatif alors que la personne présente la caractéristique à l'étude. «Un moins (1 -)» chacun de ces paramètres reflètent respectivement la spécificité et la sensibilité des méthodes de classification du test (ou de l'enquête).

Figure A. Classification croisée des résultats des tests 1 et 2				
Résultat du test 1 (Enquête initiale)	Résultat du test 2 (Réinterview)	Total		
		Positif	Négatif	Total
		Cellule 1	Cellule 2	
		Cellule 3	Cellule 4	
		n_1	n_2	$n_{..}$

Figure A. Classification croisée des résultats des tests 1 et 2

Si l'on présume que les erreurs dans le premier et le deuxième tests sont indépendantes l'une de l'autre (comme tenu de la situation vraie), les probabilités prévues, représentées par P_{ij}^g , qui sont associées aux fréquences par cellule illustrées à la figure A pour une sous-population donnée g , se définissent comme suit:

$$\text{Cellule 1 } P_{g11}^g = \pi_g^g (1 - \beta_{1,g}^g) (1 - \beta_{2,g}^g) + (1 - \pi_g^g) (\alpha_{1,g}^g \alpha_{2,g}^g)$$

$$\text{Cellule 2 } P_{g21}^g = \pi_g^g (\beta_{1,g}^g) (1 - \beta_{2,g}^g) + (1 - \pi_g^g) (\alpha_{1,g}^g) (\alpha_{2,g}^g)$$

$$\text{Cellule 3 } P_{g12}^g = \pi_g^g (1 - \beta_{1,g}^g) \beta_{2,g}^g + (1 - \pi_g^g) (\alpha_{1,g}^g) (1 - \alpha_{2,g}^g)$$

$$\text{Cellule 4 } P_{g22}^g = \pi_g^g (\beta_{1,g}^g) \beta_{2,g}^g + (1 - \pi_g^g) (1 - \alpha_{1,g}^g) (1 - \alpha_{2,g}^g). \quad (\text{A.1})$$

À partir de (A.1), nous remarquons que nous avons au total cinq paramètres mais seulement trois entrées indépendantes (ou degrés de liberté) pour faire les estimations. Le nombre de paramètres doit donc être réduit.

Pour réduire les paramètres, Hui et Walter posent d'abord comme hypothèse que la proportion de cas qui présentent la caractéristique diffère d'une sous-population à une autre et donc que $\pi_1 \neq \pi_2$. Deuxièmement, il faut que les taux d'erreur pour chaque test soient identiques dans les deux sous-populations, mais les taux d'erreur associés aux deux tests peuvent différer. Pour les deux sous-populations, ceci suppose que, dans (A.1), $\beta_r = \beta_{r,1} = \beta_{r,2}$ et $\alpha_r = \alpha_{r,1} = \alpha_{r,2}$ et que $\beta_1 \neq \beta_2$, et $\alpha_1 \neq \alpha_2$. Dans de telles conditions, le nombre de paramètres est réduit à six (deux taux de prévalence, un pour chaque sous-population, et deux taux d'erreur pour chaque test, 1 et 2). Comme les deux tableaux 2×2 contiennent six degrés de liberté, il est possible de faire des estimations. À noter que, si $\pi_1 = \pi_2$, et que les taux d'erreur sont les mêmes dans les deux sous-populations, alors les probabilités dans (A.1) seraient les mêmes pour les deux sous-populations, de sorte que nous n'aurions en fait qu'un tableau et donc qu'il serait impossible de faire des estimations. Selon le modèle de Hui et Walter (1986), les estimations par les moindres carrés non linéaires généralisés peuvent être calculées en utilisant l'algorithme de Gauss-Newton tiré de la méthode de régression non linéaire SAS. Avec cette méthode, il est possible d'exprimer les fréquences observées, n_{ij}^g , en fonction de la taille totale de l'échantillon, $n_{..}$, et de les multiplier par les probabilités dans l'expression (A.1). Hui et Walter (1986) présentent également les estimateurs en forme analytique fermée dans (A.2) en fonction des probabilités observées par cellule, lesquelles sont représentées par P_{gij}^g .

$$\hat{\alpha}_r = \frac{2E_r}{(P_{r1.}P_{r.1} - P_{r.1}P_{r1.} + P_{211} - P_{111} + D)} \quad \hat{\beta}_r = \frac{2E_r}{(P_{r.2}P_{r2.} - P_{r2.}P_{r.2} + P_{122} - P_{222} + D)}$$

où,

$$r = 2 \text{ if } r = 1, \quad r = 1 \text{ if } r = 2$$

$$P_{gij}^g = \sum_{i=1}^2 \sum_{j=1}^2 P_{gij}^g, \quad P_{g.}^g = \sum_{j=1}^2 P_{gij}^g;$$

$$\pi_g^g = \frac{1}{2} + \frac{2D}{[P_{g1.}^g(P_{1.1} - P_{2.1}) + P_{g2.}^g(P_{1.1} - P_{2.1}) + P_{211} - P_{111}]}$$

où,

$$D = \pm [(P_{11.}P_{21.} - P_{1.1}P_{21.} + P_{111} - P_{211})^2 - 4(P_{11.} - P_{21.})(P_{111}P_{2.1} - P_{211}P_{1.1})]^{\frac{1}{2}}$$

différences de légères à modérées entre les taux d'erreur chez les hommes et les femmes, en s'appuyant sur l'hypothèse de données conciliées parfaites. Il convient cependant de préciser que cette dernière hypothèse est remise en question. Prenons par exemple l'estimation de β_{121} , c'est-à-dire la probabilité qu'une personne en chômage soit classée parmi les personnes actives lors de l'enquête initiale. À partir du tableau 3, nous estimons cette valeur selon l'hypothèse d'une réinterview conciliée sans biais, en divisant n_{21} par n_1 (332/17 681 = 0,0188), où n_j a été défini précédemment et où j correspond maintenant à la situation d'activité selon la réinterview conciliée. En utilisant la valeur probable de ces deux fréquences obtenues à la section 2, nous pouvons représenter comme suit l'expression de la valeur probable de l'estimation pour de gros échantillons:

$$E(n_{21}/n_1) = \frac{\pi_1 \beta_{121} (1 - \beta_{221} - \beta_{231}) + \pi_2 (1 - \beta_{112} - \beta_{122}) \beta_{123} \beta_{213}}{\pi_1 (1 - \beta_{221} - \beta_{231}) + \pi_2 \beta_{212} + (1 - \pi_1 - \pi_2) \beta_{213}} - 1 \left[\frac{\pi_1 (1 - \beta_{221} - \beta_{231})}{\pi_1 (1 - \beta_{221} - \beta_{231}) + \pi_2 \beta_{212} + (1 - \pi_1 - \pi_2) \beta_{213}} - 1 \right] \left[\frac{\pi_1 (1 - \beta_{112} - \beta_{122}) \beta_{123} \beta_{213}}{\pi_1 (1 - \beta_{221} - \beta_{231}) + \pi_2 \beta_{212} + (1 - \pi_1 - \pi_2) \beta_{213}} + \frac{\pi_2 (1 - \beta_{112} - \beta_{122}) \beta_{123} \beta_{213}}{\pi_1 (1 - \beta_{221} - \beta_{231}) + \pi_2 \beta_{212} + (1 - \pi_1 - \pi_2) \beta_{213}} \right]. \quad (1)$$

De (1) il s'ensuit que, si le taux d'erreur dans la réinterview conciliée, β_{2ij} , est égal à zéro, alors cet estimateur est sans biais. Si toutefois la réinterview conciliée ne donne pas des résultats parfaits, le biais dans l'estimateur dépend alors des taux de prévalence dans la population à l'étude. Par conséquent, si les taux d'erreur réels dans l'enquête initiale sont en fait égaux dans les deux sous-populations étudiées et que les classifications obtenues par la conciliation ne sont pas parfaites, alors les taux d'erreur estimés dans l'enquête initiale seront différents pour les deux populations. On ne peut donc pas utiliser les similitudes ou les différences entre les taux d'erreur estimés pour les hommes et les femmes, observées lors d'études antérieures, pour confirmer ou infirmer les hypothèses présentées ici.

Nous avons également fait une analyse de sensibilité de la méthode de Hui et Walter (1986) appliquée à des réponses dichotomiques (Sincclair 1994) et constaté que la méthode est sensible, dans certains cas, à un manquement à l'hypothèse sur les taux d'erreur égaux, mais qu'elle est très robuste dans d'autres circonstances. Il faudra donc poursuivre les recherches afin de mettre au point des techniques de réinterview et d'analyse qui permettent d'assouplir les hypothèses restrictives actuellement requises pour l'analyse des données de réinterview.

Il est à noter que Chua et Fuller (1987) ont eux aussi obtenu des estimations des erreurs de classification à trois

résultats, à partir des données de réinterview correspondant à 25 % de l'échantillon de réinterview de la CPS, pour la période de 1977 à 1980. Leurs résultats, qui sont analogues aux nôtres, révèlent que les taux d'erreur les plus élevés sont associés à la classification des personnes véritablement en chômage. Poterba et Summers (1995), de même que Singh et Rao (1995), ont eux aussi constaté que ce groupe était le plus difficile à classer. Comme tous les modèles examinés indiquent que le taux global d'erreurs de classification des personnes en chômage est d'environ 20 %, les futures réinterviews devraient chercher à déterminer pourquoi ces taux sont aussi élevés et ainsi, nous l'espérons, contribuer à améliorer l'enquête.

Les estimations «corrigées», présentées au tableau 4, pourraient servir notamment à une analyse de sensibilité des études (par ex. Abowd et Zellner 1985; Poterba et Summers 1995) sur les mouvements bruts et sur la dynamique du marché du travail qui présument que l'interview conciliée était parfaite. Ceci équivalait pour eux à utiliser les estimations à l'avant-dernière colonne du tableau 3. De même, les estimations des erreurs de classification pourraient être introduites dans les méthodes utilisées pour estimer les modèles probit et logit avec les variables des réponses mal classées (Hausman et Scott-Morton 1994), ainsi que pour le développement de techniques statistiques formelles pour les données d'enquête (Rao et Thomas 1991). Il convient en terminant d'insister sur le fait que toutes les estimations corrigées en fonction des erreurs de classification n'en sont encore qu'au stade de la recherche et que les estimations des taux d'erreur ne sont pas encore suffisamment exactes pour corriger les données de l'enquête régulière, en particulier du fait qu'un nouveau questionnaire et que de nouvelles techniques d'interview ont été introduites en janvier 1994 (Bureau of Labor Statistics 1993).

REMERCIEMENTS

Les auteurs aimeraient remercier Irv Schreiner du U.S. Bureau of the Census pour les données qui ont servi à la présente étude et pour ses précieux conseils. Nous voulons également remercier John Thompson, Henry Wolman et Jon Clark, également du U.S. Bureau of the Census, de même que les examinateurs et le rédacteur adjoint, pour leurs nombreux commentaires utiles. Cette recherche a été financée en partie par une subvention de la National Science Foundation.

ANNEXE TECHNIQUE A

Examen de la méthode de Hui et Walter

La méthode de Hui et Walter a été conçue pour évaluer les tests de diagnostic. L'avantage de cette technique est qu'elle permet au chercheur de mesurer le taux d'erreur dans une épreuve donnée, sans avoir à utiliser une épreuve

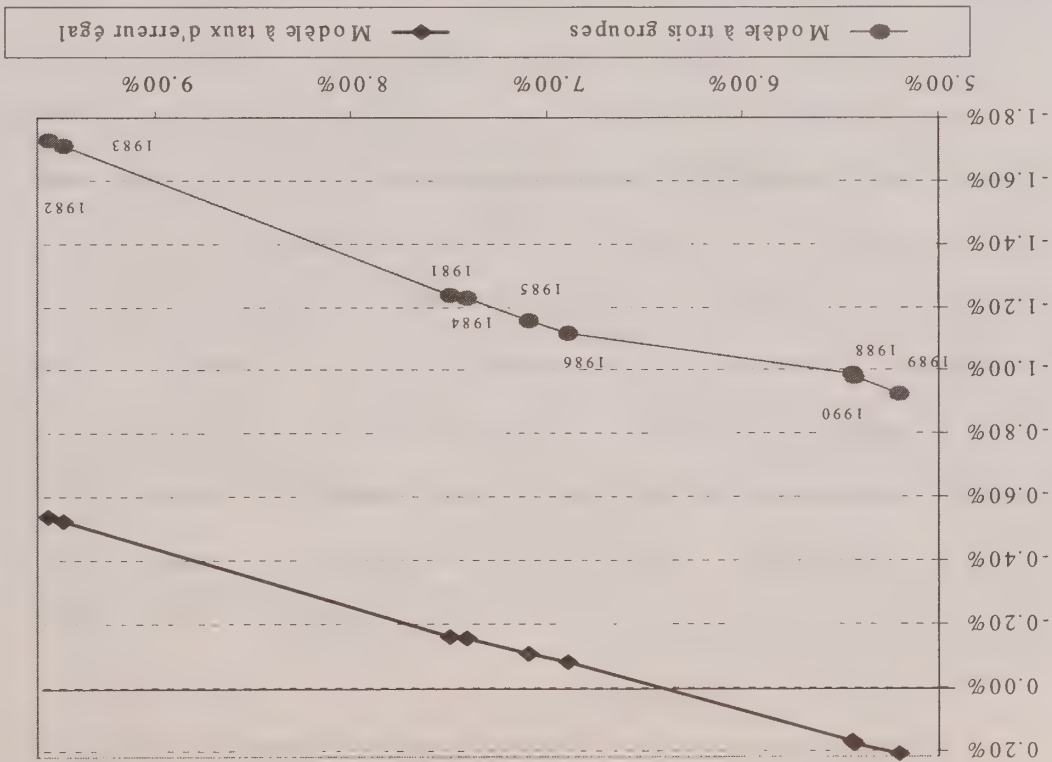


Figure 2. Comparaison entre le biais dans les taux de chômage publiés, calculé par le modèle à taux d'erreur égal et le modèle à trois groupes

tandis que le biais de surestimation peut être significatif lorsque le taux de chômage est inférieur à 5 %.

Les résultats du modèle à trois groupes laissent croire

que les taux de chômage publiés sont sous-estimés. Si cette

conclusion s'avère exacte, alors ces résultats indiquent que

le biais durant les années de faible taux de chômage est

malgré tout d'environ -0,7 % et qu'il peut atteindre jusqu'à

-1,7 % durant les années de chômage élevé. Ces résultats

tranchent avec ceux obtenus par le modèle à taux d'erreur

égal.

Le fait que l'importance et la direction du biais dans le

taux de chômage publié varient durant le cycle économique

pourrait influencer sur l'utilisation de ce taux dans les études

sur le «taux naturel» de chômage et sur la relation entre

inflation et chômage. En termes plus précis, nos résultats

indiquent que l'intervalle à l'intérieur duquel se situe le

taux réel de chômage durant le cycle économique est

supérieur à l'intervalle des taux déclarés (voir tableau 4).

Hughes et Perlman (1984) ont passé en revue la documen-

tation sur le taux «naturel» de chômage et sur la relation

entre inflation et chômage, ainsi que sur le rôle de la théorie

de recherche d'emploi pour expliquer pourquoi le chômage

n'est pas aussi faible que l'on croit, en période de «plein»

emploi. McKenna (1985) présente une étude plus poussée

sur la théorie de la recherche d'emploi et sur ses liens avec

la durée du chômage et le caractère volontaire du chômage.

Déterminer quel modèle est à la base des taux d'erreurs de

classification dans la CPS a d'importantes répercussions

6. DISCUSSION

économiques. En effet, si c'est le modèle à taux d'erreur égal qui est exact, alors le taux publié serait légèrement surestimé en périodes de faible chômage et il y aurait moins de chômage véritable à expliquer par les théories de recherche d'emploi et autres théories connexes. Par contre, si c'est le modèle à trois groupes qui est exact alors, même en périodes de chômage peu élevé, le nombre réel de chômeurs serait en fait plus élevé.

Nous avons présenté dans ce document une autre méthode pour estimer les taux d'erreur dans la CPS. Notre étude diffère des travaux antérieurs, en ce que nous suivons la méthode de Hui et Walter (1980) pour estimer les taux d'erreur, en presumant que le taux d'erreur est le même chez les hommes et les femmes et que les erreurs dans l'enquête initiale sont indépendantes de celles qui surviennent lors de la réinterview non conciliée. Bien qu'il pourrait y avoir une légère corrélation entre les erreurs, l'hypothèse d'indépendance est courante pour une analyse de données de ce type (voir Bailar 1968, Chua et Fuller 1987 et Singh et Rao 1995). On trouve une discussion du biais de la méthode de H et W avec erreurs dépendantes dans Vasek (1995). Quant à l'hypothèse d'un taux d'erreur égal entre les deux sous-populations, plusieurs auteurs cités ici (par ex., Poterba et Summers 1986) ont relevé des

Taux d'erreur estimés dans les données de la CPS initiale, pour trois catégories du taux de chômage

Paramètre du taux d'erreur	Description	Estimations des taux d'erreur									
Classes parmi	Situation d'activité véritable	Modèle au tableau 1 : taux d'erreur constant durant toutes les années	Années à faible taux 1990,1989 et 1988			Années à taux modéré 1981, 1984-1986			Années à taux élevé 1982, 1983		
			Est.	E-T	Est.	E-T	Est.	E-T	Est.	E-T	
β_{121}	Occupés	Chômeurs	0,407	0,0189	0,0635	0,1061	0,1113	0,1258	0,0974	0,0717	
β_{131}	Inactifs	Chômeurs	0,1196	0,0146	0,1680	0,0538	0,1000	0,0246	0,1084	0,0221	
β_{112}	Chômeurs	Occupés	0,0049	0,0012	0,0000	0,0047	0,0000	0,0098	0,0000	0,0069	
β_{132}	Inactifs	Occupés	0,0100	0,0015	0,0080	0,0038	0,0096	0,0025	0,0096	0,0031	
β_{113}	Chômeurs	Inactifs	0,0110	0,0015	0,0096	0,0040	0,0109	0,0024	0,0103	0,0029	
β_{123}	Occupés	Inactifs	0,0205	0,0025	0,0187	0,0065	0,0202	0,0034	0,0227	0,0044	

Tableau 6

Répercussions des estimations des taux d'erreur, selon le modèle à trois groupes

Année y	Taux de chômage publié par le BLS	Prob en chômage d'après ^{hrc} Modèle classé en chômage –	Estimation corrigée du taux de chômage publié par le BLS	Modèle initial à taux d'erreur égal	Modèle à trois groupes	Modèle initial à taux d'erreur égal	Modèle à trois groupes	Erreur-type estimée de la différence – méthode à trois groupes
1990	5,44 %	0,9124	5,27 %	6,43 %	0,17 %	-0,99 %	1,40 %	
1989	5,20 %	0,9088	4,99 %	6,12 %	0,21 %	-0,93 %	1,35 %	
1988	5,43 %	0,9105	5,25 %	6,41 %	0,18 %	-0,98 %	1,41 %	
1986	6,89 %	0,9170	6,97 %	8,01 %	-0,08 %	-1,12 %	2,35 %	
1985	7,09 %	0,9178	7,20 %	8,25 %	-0,11 %	-1,16 %	2,42 %	
1984	7,41 %	0,9199	7,56 %	8,64 %	-0,15 %	-1,23 %	2,53 %	
1983	9,47 %	0,9400	9,99 %	11,18 %	-0,52 %	-1,71 %	2,05 %	
1982	9,54 %	0,9404	10,08 %	11,27 %	-0,54 %	-1,73 %	2,08 %	
1981	7,50 %	0,9191	7,66 %	8,74 %	-0,16 %	-1,24 %	2,56 %	

existent entre les estimations du biais obtenues par les deux méthodes, il est difficile de déterminer l'importance du biais. Malheureusement, les estimations sont sensibles à la spécification du modèle, en raison de la petite taille de l'échantillon pour la réinterview non conciliée; ceci se reflète dans les erreurs-types élevées qui sont associées aux estimations du taux d'erreur et, partant, aux estimations du biais.

5. INCIDENCE DES ESTIMATIONS CORRIGÉES

Les résultats illustrés aux figures 1 et 2 montrent que toutes les méthodes utilisées pour corriger le taux de chômage en fonction des erreurs de classification indiquent que l'importance du biais dans le taux publié varie tout au long du cycle économique. Compte tenu des différences qui

Notre approche, qui présume que le taux d'erreur demeure constant durant toute la période, laisse croire que le biais dans les estimations de l'enquête est faible durant les années où le taux de chômage se situe entre 5,5 % et 7,5 %. Avec ce modèle, le taux de chômage publié semble sans biais lorsque le taux de chômage vrai se situe autour de 6,3 %, alors qu'il est sous-estimé lorsque le taux réel est supérieur à ce seuil et qu'il est surestimé lorsqu'il est inférieur. Par ailleurs, le biais dû à la sous-estimation devient assez apparent lorsque le taux de chômage atteint 9 %.

5. INCIDENCE DES ESTIMATIONS CORRIGÉES

La figure 2 présente un graphique qui illustre le biais dans les taux de chômage calculés par le modèle à trois groupes et, pour fins de comparaison, par le modèle à taux d'erreur égal. Les résultats à la figure 2 sont très intéressants. Ils montrent ainsi que, même si l'effet cyclique est toujours présent, le biais estimé est ramené vers le bas et présente un biais négatif constant pendant toute la durée du cycle économique.

Les résultats illustrés aux figures 1 et 2 montrent que toutes les méthodes utilisées pour corriger le taux de chômage en fonction des erreurs de classification induisent une l'importance du biais dans le taux publié varie tout au long du cycle économique. Compte tenu des différences qui

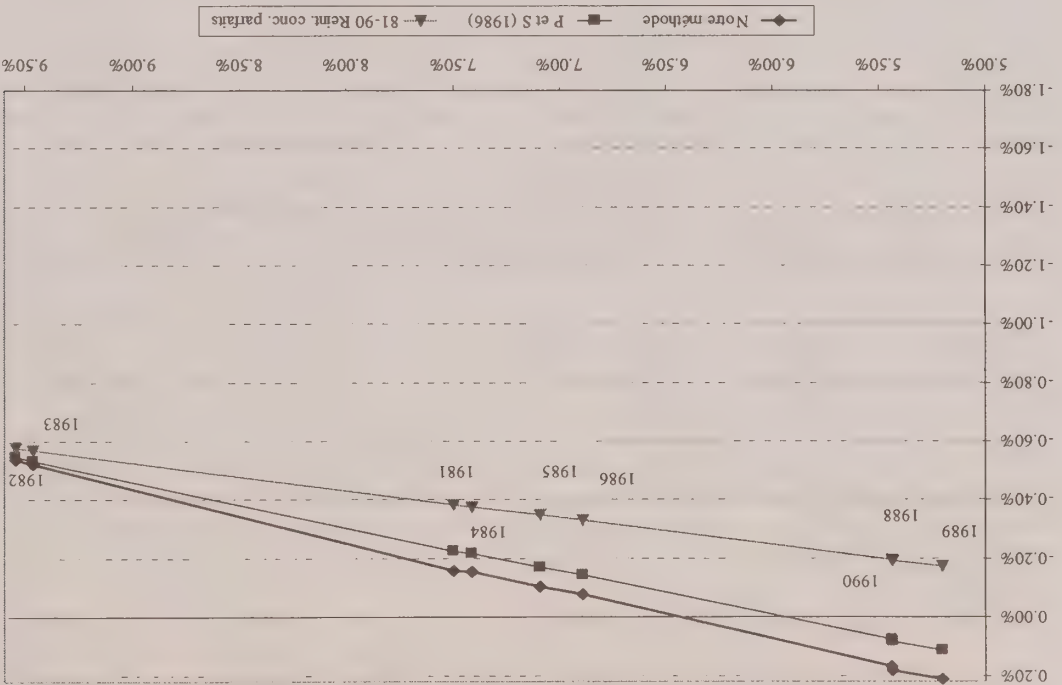


Figure 1. Comparaison entre le biais dans les taux de chômage publiés, selon les trois méthodes de calcul

Par ailleurs, même si les résultats du test du rapport des vraiesemblances indiquaient des taux d'erreur constants durant toute la période, les examinateurs ont proposé de pousser plus loin l'analyse de cette hypothèse. Nous avons donc réparti chacune des neuf années d'enquête en trois groupes, en fonction du taux de chômage publié pour chacune de ces années. Ainsi, les années d'enquête 1990, 1989 et 1988 ont été classées dans le groupe à faible taux de chômage, les taux déclarés pour ces années variant de 5,20 % à 5,44 %. À l'autre extrémité, les années 1982 et 1983 ont été classées dans la catégorie à taux élevé de chômage, avec des taux respectifs de 9,54 % et 9,47 %. Enfin, les autres années ont été classées dans le groupe à taux de chômage modéré, avec des taux variant de 6,89 %

Dans les études sur les tests de dépistage (Gastwirth 1987), la fraction des classifications positives qui sont exactes, appelée valeur prédictive d'un test positif, varie directement en fonction de la prévalence de la caractéristique. C'est ce qui explique que des tests de diagnostic assez précis peuvent donner des taux d'erreur inacceptables lorsqu'ils sont utilisés dans des populations où la prévalence de la maladie à l'étude est faible. L'analogie entre cette mesure et notre étude se situe dans la proportion des personnes qui sont classées parmi les chômeurs et qui sont véritablement en chômage. Cette proportion est indiquée à la troisième colonne, du tableau 4. Bien que l'intervalle des taux de chômage publiés soit assez restreint, on peut néanmoins voir une relation similaire avec le taux

Pour évaluer la sensibilité des estimations corrigées du taux de chômage, au tableau 4, nous les avons recalculées en utilisant les taux d'erreur obtenus par le modèle à trois groupes. Ces résultats sont présentés au tableau 6, lequel indique également l'erreur-type des estimations du taux de chômage, qui varie de près de 1,4 % à un sommet d'environ 2,6 %.

Dans l'ensemble, les estimations des taux d'erreur, pour les trois catégories du taux de chômage, semblent similaires. Cependant, comme les erreurs-types des taux d'erreur estimés sont assez élevées, un test formel d'homogénéité n'aurait pas une puissance suffisante pour déceler toute variation dans le taux d'erreur, sur les trois périodes. Dans l'ensemble, les estimations des taux d'erreur, pour les trois catégories du taux de chômage, semblent similaires. Cependant, comme les erreurs-types des taux d'erreur estimés sont assez élevées, un test formel d'homogénéité n'aurait pas une puissance suffisante pour déceler toute variation dans le taux d'erreur, sur les trois périodes.

Nous avons fait un test du rapport des vraiesemblances, pour vérifier l'hypothèse selon laquelle le taux d'erreur était égal à l'intérieur de chacun de ces trois groupes, en établissant une comparaison avec le modèle initial sur neuf années. La valeur du ratio de vraiesemblance, $-2 \log \lambda$, avec 72 degrés de liberté (144 paramètres dans le modèle complet moins 72 paramètres dans le modèle à trois groupes) a été de 69,25, pour une valeur de p de 0,5697.

Tableau 1

Taux d'erreur estimés dans les données de la CPS initiale

Paramètre du taux d'erreur	Description	Valeur estimée β_{ij}	Erreur-type estimée
	Classe parmi	Situation véritable	Notre méthode
β_{121}	Occupés	Chômeurs	0,0407
β_{131}	Inactifs	Chômeurs	0,1196
β_{112}	Chômeurs	Occupés	0,0049
β_{132}	Inactifs	Occupés	0,0100
β_{113}	Chômeurs	Inactifs	0,0110
β_{123}	Occupés	Inactifs	0,0205
			0,0116
			0,0064
			0,0098
			0,0017
			0,0838
			0,0188

Réint. conc.
partielle

Tableau 2

Taux d'erreur estimés à partir de la réinterview non conciliée de la CPS

Paramètre du taux d'erreur	Description	Valeur estimée	Notre méthode	Erreur-type estimée
	Classe parmi	Situation véritable	Notre méthode	
β_{121}	Occupés	Chômeurs	0,0333	0,01772
β_{131}	Inactifs	Chômeurs	0,1128	0,01360
β_{112}	Chômeurs	Occupés	0,0057	0,00135
β_{132}	Inactifs	Occupés	0,0145	0,00160
β_{113}	Chômeurs	Inactifs	0,0157	0,00171
β_{123}	Occupés	Inactifs	0,0248	0,00238

 β_{2ij} Totalisations croisées entre les réponses agrégées de l'enquête initiale et celles de la réinterview conciliée, 1981-1990
Données de réinterview conciliées correspondant à 75 % de l'échantillon de réinterview de la CPS

Résultat de l'enquête	Chômeurs	Occupés	Inactifs	Total
CPS initiale	Chômeurs	Occupés	Inactifs	Total
15 868	372	480	16 720	
332	213 987	744	215 063	
1 481	2 123	138 077	141 681	
17 681	215 482	139 301	373 464	

Réinterview conciliée

Tableau 4

Incidence des estimations des taux d'erreur

Année y	Taux de chômage publié par le BLS	Prob. en chômage, au nombre de classes comme chômeurs	Notre méthode	Poterna et Sumners (1986)	Données conciliées (1981-1990)	Notre méthode	Erreur-type estimée de la différence
1990	5,44 %	0,8135	5,27 %	5,36 %	5,63 %	0,17 %	0,27 %
1989	5,20 %	0,8052	4,99 %	5,09 %	5,37 %	0,21 %	0,26 %
1988	5,43 %	0,8113	5,25 %	5,35 %	5,62 %	0,18 %	0,27 %
1986	6,89 %	0,8503	6,97 %	7,04 %	7,22 %	-0,08 %	0,33 %
1985	7,09 %	0,8531	7,20 %	7,27 %	7,44 %	-0,11 %	0,34 %
1984	7,41 %	0,8581	7,56 %	7,63 %	7,79 %	-0,15 %	0,36 %
1983	9,47 %	0,8894	9,99 %	10,00 %	10,04 %	-0,52 %	0,48 %
1982	9,54 %	0,8902	10,08 %	10,09 %	10,12 %	-0,54 %	0,49 %
1981	7,50 %	0,8581	7,66 %	7,72 %	7,88 %	-0,16 %	0,36 %

4. ANALYSE DES DONNÉES ET RÉSULTATS

Le tableau présente les données agrégées par année et par sexe, de sorte que les taux d'erreur calculés à partir de ce tableau peuvent être comparés à notre modèle. En utilisant les données figurant à la colonne «situation véritable», nous calculons une estimation du taux d'erreur. À titre d'exemple, l'estimation de β_{121}^* , qui représente la probabilité qu'une personne en chômage soit classée parmi les personnes actives lors de l'enquête initiale, est de $332/17\,681 = 0,0188$. Ces taux d'erreur sont présentés au tableau 1, pour illustrer dans quelle mesure les taux d'erreur estimés par notre méthode basée sur les données non conciliées diffèrent de ceux obtenus en présupposant que la réinterview conciliée donne des résultats parfaits.

Le tableau 1 présente également les estimations des taux d'erreur pour l'enquête initiale, calculés par Poterba et Summers (1986) à partir des données de la réinterview (combinaisons pour les deux sexes), pour la première moitié de 1981. Poterba et Summers (1986) s'appuient sur des données conciliées et non conciliées pour estimer les taux d'erreur. Ces auteurs présument que, dans l'échantillon concilié, les intervieweurs utilisent les données de l'enquête initiale pour influencer la réponse au moment de la première réinterview. Ils supposent donc qu'il y a production d'une valeur conciliée uniquement pour la proportion des répondants pour qui il devrait y avoir divergence entre les résultats de l'enquête initiale et ceux de la première réinterview. Par ailleurs, lorsqu'ils obtiennent une valeur conciliée, Poterba et Summer (1986) présument que celle-ci est exempte d'erreur. À partir de ces hypothèses ils utilisent l'échantillon non concilié pour estimer la fréquence de l'erreur et les données conciliées pour fournir de l'information sur la véritable situation d'activité. En résumé, la méthode de Poterba et Summers (1986) et la réinterview conciliée estiment toutes deux que les données conciliées de la réinterview sont parfaites.

Le tableau 4 présente les taux de chômage annuels publiés par le BLS, hommes et femmes combinés, et les comparer aux taux de chômage corrigés d'après: (1) nos estimations du taux d'erreur, (2) les taux d'erreur calculés par Poterba et Summers (1986) et (3) les taux d'erreur obtenus en présupposant que les données conciliées de la réinterview sont parfaites. Lorsqu'on répartit les résultats du tableau 4 en fonction de la valeur du taux de chômage publié par le BLS, il s'en dégage une tendance évidente, quant au biais dans les estimations initiales de la CPS. La figure 1 montre ainsi que les chiffres publiés ont tendance à surestimer le taux de chômage réel durant les années où le chômage est faible (1989, 1988 et 1990) et, inversement, à sous-estimer durant les années où le chômage est élevé (1982-1983). On remarque en outre qu'un biais vers le haut est associé à notre méthode, comparativement aux deux autres. Enfin, les trois méthodes indiquent un effet cyclique, qui est le plus faible lorsqu'on présuppose que la réinterview conciliée est parfaite.

La première étape dans la préparation de nos estimations finales a été d'obtenir les estimations de paramètres pour chacune des neuf tables de données annuelles, en utilisant la méthode de régression non linéaire SAS avec la méthode des moindres carrés généralisés de Gauss-Newton. Comme les méthodes de réinterview sont demeurent inchangées durant cette période, nous avons décidé de vérifier l'hypothèse selon laquelle les taux d'erreur étaient égaux pendant toutes les années à l'étude, c'est-à-dire que $\beta_{ygrj}^* = \beta_{y'grj}^*$ pour toutes les années $y \neq y'$. Si on combine maintenant cette hypothèse à l'hypothèse de base voulant que les taux d'erreur pour les hommes et les femmes soient égaux, c'est-à-dire que $\beta_{y1rj}^* = \beta_{y2rj}^*$, alors cela suppose que $\beta_{y'grj}^* = \beta_{y'g'rj}^*$ pour toutes les valeurs de $y \neq y'$ et $g \neq g'$. À partir des deux séries de résultats, nous avons effectué un test du rapport des vraisemblances, en présupposant que l'échantillon de la réinterview est un échantillon aléatoire simple de la population, utilisé pour vérifier l'hypothèse voulant que chaque taux d'erreur est égal pour toutes les années. La valeur du ratio de vraisemblance, $-2 \log \lambda$ avec 96 degrés de liberté (144 paramètres dans le modèle intégral, moins 48 paramètres dans le modèle réduit), est de 84,06 pour une valeur de p de 0,8027. Les données sont donc cohérentes avec le modèle réduit, ce qui nous permet d'utiliser les estimations du modèle réduit et de simplifier la notation. Aussi, à partir de maintenant, nous utiliserons β_{ygrj}^* pour représenter β_{ygrj}^* pour toutes les valeurs de g et y .

Les estimations du taux d'erreur pour l'enquête initiale et la réinterview non conciliée sont présentées respectivement aux tableaux 1 et 2, avec leurs erreurs-types estimées. Nous remarquons que les taux d'erreur estimés pour la réinterview (tableau 2) sont similaires aux taux d'erreur correspondants de l'enquête initiale. Cette similitude indique que la réinterview non conciliée, qui est menée par le U.S. Bureau of the Census, est une répétition efficace. Les estimations du taux d'erreur montrent en outre que les techniques d'enquête utilisées pour la CPS permettent de classer de façon assez exacte les personnes occupées et inactives. En revanche, ces techniques ne donnent pas d'aussi bons résultats pour ce qui est du classement des personnes en chômage, car la proportion des chômeurs qui sont en fait classés parmi les chômeurs, $(1 - \beta_{121}^* - \beta_{131}^*)$, n'est que de 0,8397. Pour fins de comparaison, nous avons fait une analyse des données conciliées de la réinterview — lesquelles correspondent à 75 % de l'échantillon intégral — et ce pour la même période (1981 à 1990), en présupposant que les réponses conciliées étaient exemptes d'erreur. Nous avons la aussi créé un tableau 3×3 , en y indiquant le nombre de personnes classées dans chacune des trois catégories d'activité au moment de l'enquête initiale, en fonction du nombre de personnes classées dans chaque catégorie, selon la réinterview conciliée. Ces données sont indiquées au tableau 3.

réinterview non conciliée, et qui sont liées à la situation d'activité réelle du répondant, sont indépendantes. Pour appliquer la méthode de Hui et Walter, il nous faut deux sous-populations dans lesquelles les taux de prévalence diffèrent. Comme il est un fait connu que les taux d'activité des hommes et des femmes diffèrent, nous avons choisi d'utiliser ces deux groupes. Nous devons également présumer que le taux d'erreurs de classification est égal dans les deux sous-populations, hommes et femmes, c'est-à-dire que $\beta_{y1r1j} = \beta_{y2r1j}$. À ce stade-ci, nous présumons également que les taux d'erreurs de classification lors de l'enquête initiale et de la réinterview non conciliée peuvent être différents et qu'ils peuvent aussi varier d'une année à l'autre. Compte tenu de cette réduction, nous obtenons pour les deux sous-populations, pour une année donnée, 12 paramètres du taux d'erreur et 4 taux de prévalence, pour un total de 16 paramètres. Comme deux tableaux 3×3 contiennent au total 16 degrés de liberté, il est possible de faire une estimation. Pour l'étude présentée ici, nous avons analysé les données de l'échantillon de réinterview non conciliée de la CPS, pour la période allant de 1981 à 1990. Les données annuelles complètes pour 1987 et d'autres données plus récentes n'ont pu être obtenues du U.S. Bureau of the Census.

Le Bureau of Labor Statistics (BLS) publie régulièrement des estimations du taux de chômage, à partir des données de la CPS (voir Bureau of Labor Statistics 1992). Comme la réinterview est un sous-échantillon de l'échantillon intégral utilisé pour la CPS, les estimations du taux de chômage établies à partir de l'échantillon de réinterview différeront des données publiées par le BLS. En outre, des techniques de traitement des données sont appliquées à l'échantillon intégral de la CPS, mais non aux données de la réinterview. À titre d'exemple, l'échantillon complet de la CPS est pondéré en fonction de la probabilité de sélection, et des facteurs de correction pour la non-réponse sont appliqués aux données. En raison de ces différences, les prévalences estimées par notre modèle, à partir uniquement des données de la réinterview, ne peuvent être comparées directement aux données publiées par le BLS. Nous avons utilisé les données de réinterview principalement pour estimer les taux d'erreur dans l'enquête initiale. De plus, tout au long du présent document, nous considérons que les données non conciliées de la réinterview correspondent à un échantillon aléatoire simple de la population aux fins des analyses et des tests d'hypothèses. À partir de ces estimations du taux d'erreur, nous estimons les taux de chômage corrigés du Bureau of Labor Statistics (BLS), le terme «corrigé» signifiant ici que les chiffres publiés ont été modifiés pour tenir compte des erreurs de classification au moment de l'enquête. La formule qui est utilisée pour estimer le taux de chômage vrai, en fonction des prévalences publiées par le BLS d'après l'échantillon complet de la CPS et des taux d'erreurs de classification estimés à partir des données non conciliées de la réinterview, est indiquée en annexe.

tableau 3×3 pour une année donnée et une sous-population donnée ne compte que 8 fréquences indépendantes, ou degrés de liberté. Le modèle compte donc trop de paramètres et il faut réduire leur nombre; nous utilisons pour ce faire le paradigme de Hui et Walter.

3. APPLICATION DU MODÈLE ET DU PROGRAMME DE RÉINTERVIEW DE LA CPS

Le programme de réinterview prévu dans le cadre de la Current Population Survey (CPS) du U.S. Bureau of the Census (U.S. Bureau of the Census 1963) est effectué deux semaines environ après l'enquête initiale, pour mesurer l'erreur de réponse et évaluer la performance de l'intervieweur. Le plan d'échantillonnage pour la réinterview consiste en un échantillon aléatoire auto-pondéré des ménages (Levy et Lemeshow 1980), choisi parmi les ménages affectés à l'intervieweur évalué. La taille de l'échantillon correspond à environ 1/18 de l'échantillon mensuel intégral de la CPS, lequel compte entre 50 000 et 60 000 ménages. Deux méthodes de réinterview sont utilisées. Dans un premier temps, une proportion allant des trois quarts aux quatre cinquièmes de l'échantillon participe à une étude sur le biais dans les réponses, pour laquelle on procède à une première réinterview; une fois cette réinterview terminée, l'intervieweur concilie les divergences entre les réponses obtenues à l'enquête initiale et celles de la première réinterview. Durant cette étude sur le biais dans les réponses, on peut donc obtenir jusqu'à deux réponses de réinterview pour un même sujet, la première étant la réponse non conciliée et la deuxième, la réponse de réinterview conciliée. Dans un deuxième temps, le reste des ménages (soit entre un cinquième et un quart de l'échantillon intégral) passent une réinterview sans qu'il y ait conciliation des résultats.

Pour l'étude sur le biais dans les réponses, l'intervieweur ne doit regarder les réponses de l'enquête initiale qu'après que la réinterview est terminée. Selon Forsman et Schreiner (1991) et Schreiner (1980), l'intervieweur modifierait la réponse obtenue lors de la première réinterview pour qu'elle concorde avec la réponse de l'enquête initiale, car ils ont constaté que l'écart entre les réponses de l'enquête initiale et celles de la première réinterview était plus grand dans l'échantillon non concilié. Sinclair (1994) et Sinclair et Gastwirth (1996) ont ensuite démontré que ces différences étaient statistiquement significatives. Le processus de conciliation crée donc une corrélation entre les réponses initiales et les réponses non conciliées de la réinterview, dans l'échantillon concilié. Aussi, avons-nous décidé de limiter notre analyse aux données de l'enquête initiale et aux données non conciliées de la réinterview. Pour notre étude, nous présumons que, dans l'échantillon non concilié, les erreurs qui proviennent de l'enquête initiale et de la

section trois. La section quatre présente les taux d'erreur obtenus, ainsi que les taux de chômage « corrigés » en fonction des erreurs de classification estimées. Une mesure de l'exacitude, dite valeur prédictive, qui est utilisée dans les ouvrages sur les tests médicaux de dépistage, est également appliquée au taux de chômage, à la section quatre. Cette mesure indique que la probabilité qu'une personne classée parmi les chômeurs lors de la CPS soit véritablement en chômage varie en fonction du taux réel de chômage.

2. LES DONNÉES ET LE MODÈLE

Les données de réinterview de la population active consistent en des réponses trinomiales tirées à la fois de l'enquête initiale et d'une réinterview menée par la suite. Ces données pour une sous-population et une année données sont résumées dans un tableau 3 × 3, où la fréquence observée des sujets est représentée par n_{ygi} .

Dans cette notation :

- y représente l'année;
- g représente la sous-population, $g = 1$ ou 2;
- i représente la situation d'activité du sujet selon l'enquête initiale: $i = 1$, chômeur; $i = 2$, occupé et $i = 3$, inactif, et
- j représente la situation d'activité du même sujet, selon la réinterview ($j = 1, 2$ et 3).

Pour chaque situation d'activité, $i = 1, 2$ et 3, le taux de prévalence vrai, pour la sous-population g et l'année y , est représenté par π_{ygi} . Tout au long du document, nous utiliserons l'expression «taux de prévalence» pour désigner la proportion de personnes dans l'une des trois situations d'activité (par ex., π_{ygi}). À noter que la fraction π_{ygi} dans la catégorie des inactifs est égale à $(1 - \pi_{ygi} - \pi_{ygi})$ et que le taux de chômage réel durant l'année y , pour la sous-population g , est égal à π_{ygi} divisé par $(\pi_{ygi} + \pi_{ygi})$.

Chaque taux de classification, β_{ygrj} , est défini comme la probabilité que le r -ième processus de collecte de données ($r = 1$ pour l'enquête initiale et $r = 2$ pour la réinterview) classe une personne de la sous-population g , durant l'année y , comme appartenant à la catégorie i , $i = 1, 2$ et 3, lorsque la personne se classe en fait dans la catégorie j . Par exemple, β_{11131} indique la probabilité que, durant la première année ($y = 1$), une personne de la première sous-population ($g = 1$) soit classée parmi la population inactive ($i = 3$), au moment de l'enquête initiale ($r = 1$), alors que cette personne est en fait en chômage ($j = 1$). On peut répartir les taux de classification en deux groupes, à savoir les taux associés à une classification exacte et ceux associés à une classification erronée. Pour chaque y , g et r , la probabilité que la méthode d'enquête r , durant l'année y , classe correctement un chômeur de la sous-population g comme étant en chômage est égale à $\beta_{ygr11} = (1 - \beta_{ygr21} - \beta_{ygr31})$. Les probabilités correspondantes pour les personnes occupées et inactives sont, respectivement,

$\beta_{ygr22} = (1 - \beta_{ygr12} - \beta_{ygr32})$ et $\beta_{ygr33} = (1 - \beta_{ygr13} - \beta_{ygr23})$. Si l'on présume d'une indépendance conditionnelle entre les taux de classification selon l'enquête initiale et la réinterview, alors les fréquences observées probables, exprimées par la notation pour chacune des neuf cellules associées à une année y donnée et à la sous-population g , sont :

$$E(n_{ygi1}) = n_{ygi} (\pi_{ygi} (1 - \beta_{ygi12} - \beta_{ygi13}) + (1 - \pi_{ygi}) \beta_{ygi13})$$

$$E(n_{ygi2}) = n_{ygi} (\pi_{ygi} (1 - \beta_{ygi12} - \beta_{ygi13}) + \pi_{ygi} \beta_{ygi23})$$

$$* (1 - \beta_{ygi21} - \beta_{ygi22}) + (1 - \pi_{ygi}) \beta_{ygi13} \beta_{ygi23})$$

$$E(n_{ygi3}) = n_{ygi} (\pi_{ygi} (1 - \beta_{ygi12} - \beta_{ygi13}) + \pi_{ygi} \beta_{ygi23} + \pi_{ygi} \beta_{ygi12} \beta_{ygi23})$$

$$+ (1 - \pi_{ygi}) \beta_{ygi13} (1 - \beta_{ygi23} - \beta_{ygi22}))$$

$$E(n_{ygi1}) = n_{ygi} (\pi_{ygi} \beta_{ygi12} (1 - \beta_{ygi21} - \beta_{ygi23})$$

$$+ \pi_{ygi} (1 - \beta_{ygi12} - \beta_{ygi13}) \beta_{ygi23} + (1 - \pi_{ygi}) \beta_{ygi13} \beta_{ygi23})$$

$$E(n_{ygi2}) = n_{ygi} (\pi_{ygi} \beta_{ygi12} \beta_{ygi21} + \pi_{ygi} (1 - \beta_{ygi12} - \beta_{ygi13})$$

$$* (1 - \beta_{ygi21} - \beta_{ygi22}) + (1 - \pi_{ygi}) \beta_{ygi13} \beta_{ygi23})$$

$$E(n_{ygi3}) = n_{ygi} (\pi_{ygi} \beta_{ygi12} \beta_{ygi21} + \pi_{ygi} (1 - \beta_{ygi12} - \beta_{ygi13}) \beta_{ygi23}$$

$$+ (1 - \pi_{ygi}) \beta_{ygi13} (1 - \beta_{ygi23} - \beta_{ygi22}))$$

$$E(n_{ygi1}) = n_{ygi} (\pi_{ygi} \beta_{ygi13} (1 - \beta_{ygi21} - \beta_{ygi23})$$

$$+ \pi_{ygi} \beta_{ygi13} \beta_{ygi21} + (1 - \pi_{ygi}) \beta_{ygi13} \beta_{ygi23})$$

$$E(n_{ygi2}) = n_{ygi} (\pi_{ygi} \beta_{ygi13} \beta_{ygi21} + \pi_{ygi} \beta_{ygi13} (1 - \beta_{ygi21} - \beta_{ygi23})$$

$$+ (1 - \pi_{ygi}) \beta_{ygi13} (1 - \beta_{ygi21} - \beta_{ygi23}))$$

$$E(n_{ygi3}) = n_{ygi} (\pi_{ygi} \beta_{ygi13} \beta_{ygi21} + \pi_{ygi} \beta_{ygi13} \beta_{ygi23}$$

$$+ (1 - \pi_{ygi}) \beta_{ygi13} (1 - \beta_{ygi21} - \beta_{ygi23}))$$

où la taille de l'échantillon pour l'année y et la sous-population g est représentée par n_{ygi} . Le modèle compte 14 paramètres (six taux d'erreur pour l'enquête initiale, $r = 1$, six taux d'erreur pour la réinterview, $r = 2$, et deux taux de prévalence uniques) pour chaque sous-population et année. Cependant, le

Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publiés

MICHAEL D. SINCLAIR et JOSEPH L. CASTWIRTH¹

RÉSUMÉ

Nous examinons dans cet article les erreurs de réponse qui surviennent lors de la Current Population Survey menée par le U.S. Bureau of the Census et évaluons l'incidence de ces erreurs sur les taux de chômage qui sont ensuite publiés par le Bureau of Labor Statistics. Les taux d'erreur sont calculés à partir des données de la réinterview, par une méthode qui se veut un prolongement de celle mise au point par Hui et Walter pour évaluer les tests de diagnostic. Contrairement aux études antérieures qui présupposaient que les données conciliaient la réinterview reflétaient la situation véritable, la méthode proposée ici estime les taux d'erreur dans les deux interviews. À partir de ces estimations, nous montrons que les erreurs de classification dans l'enquête initiale ont un effet cyclique sur les estimations du taux de chômage qui sont ensuite publiées et que le degré de sous-estimation augmente tout particulièrement lorsque le taux de chômage réel est élevé. Comme nous n'avions pas suffisamment de données pour établir une distinction entre, d'une part, un modèle présumant que les taux d'erreurs de classification sont les mêmes durant tout le cycle économique et, d'autre part, un modèle qui présume que les taux varient selon que le taux de chômage est faible, modéré ou élevé, nos conclusions doivent être considérées comme provisoires. Celles-ci indiquent néanmoins qu'il y aurait lieu d'examiner plus à fond la relation entre les modèles utilisés pour évaluer l'exactitude des tests de diagnostic et les modèles servant à mesurer les taux d'erreurs de classification dans les données d'enquête.

MOTS CLÉS : Erreurs de classification; taux de chômage; test de diagnostic; conciliation; enquête de réinterview; erreurs de réponse.

1. INTRODUCTION

Plusieurs articles, notamment ceux de Poterba et Summers (1986 et 1995) et d'Abowd et Zellner (1985), ont utilisé les données du programme de réinterview mis en place par le U.S. Bureau of the Census, pour estimer les taux d'erreurs de classification dans la Current Population Survey (CPS) et évaluer l'incidence de ces erreurs sur les estimations des changements de situation vis-à-vis de l'activité sur le marché du travail. Ces estimations des taux d'erreurs de classification étaient basées sur l'hypothèse selon laquelle une méthode particulière de réinterview, dite interview de conciliation, donne les « vraies » valeurs. Cependant, Biemer et Forsman (1992), Forsman et Schreiner (1991) et des recherches non publiées menées par le U.S. Bureau of the Census (1963) ont mis en doute cette hypothèse. Le but du présent article est de fournir des estimations des taux d'erreurs de classification dues aux erreurs de réponse lors des interviews et des réinterviews et d'en examiner l'incidence sur les taux de chômage publiés. Contrairement aux études antérieures qui ont porté sur les mouvements bruts, nous nous intéressons ici à l'exactitude des estimations sur la main-d'oeuvre. Notre approche consiste à appliquer à des classifications trinomiales le paradigme proposé par Hui et Walter (1980) pour estimer les taux d'erreur des tests pour diagnostic médical. Un des avantages de cette méthode est qu'il n'est pas nécessaire d'avoir une interview que l'on considère parfaite.

À partir de certaines hypothèses, Hui et Walter (1980) ont mis au point une méthode pour estimer les taux d'erreur associés à un nouveau test de dépistage, en utilisant une épreuve de confirmation à faible taux d'erreur inconnu. Si l'on considère ici que la réinterview équivaut à l'épreuve de confirmation et l'enquête initiale au test de dépistage, cette méthode peut être utilisée pour estimer les taux d'erreur dans l'enquête initiale et au moment de la réinterview, ainsi que les taux de prévalence de la caractéristique à l'étude. La méthode de Hui et Walter (1980) nécessite deux sous-populations dans lesquelles le taux de prévalence de la caractéristique à l'étude diffère. Même si le taux d'erreur n'est pas nécessairement identique dans les deux tests, on présume que les taux d'erreur sont égaux dans les deux sous-populations, pour chaque test. Le modèle (décrit plus en détail en annexe) présume en outre que les erreurs, qui dépendent de la véritable situation d'activité du sujet, sont indépendantes d'un test à un autre.

La méthode de Hui et Walter a été conçue pour des résultats dichotomiques et a été adaptée à l'étude des erreurs de classification des taux d'activité de la population active, par Sinclair et Gastwirth (1996). Nous élargissons ici cette approche pour l'appliquer à trois classifications: chômeurs, population occupée et population inactive, et nous évaluons l'effet des erreurs de classification sur les taux de chômage publiés. Le modèle de base est décrit à la section deux. Les données du programme de réinterview, sur lesquelles s'appuie notre modèle, sont décrites à la

¹ Michael D. Sinclair, Senior Statistician, Mathematical Policy Research, 101 Morgan Lane, Plainsboro, N.J. 08536, U.S.A.; Joseph L. Gastwirth, Professor of Statistics and Economics, George Washington University, 2201 G Street Rm. 315, Washington, D.C. 20052, U.S.A.

REMERCIEMENTS

Les auteurs tiennent à remercier Douglas Yeo ainsi que les examinateurs pour leurs judicieux commentaires.

BIBLIOGRAPHIE

KASS, G.V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, 29, 119-127.

LAVALLÉE, P. (1995). Pondération transversale des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-35.

LITTLE, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *Revue Internationale de Statistique*, 54, 137-157.

OH, H.T., et SCHEUREN, F. (1983). Weighting adjustments for unit nonresponse. Dans *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, (Éds. W.G. Madow, I. Olkin et D. Rubin). New York: Academic Press, 143-184.

RIZZO, L., KALTON, G., et BRICK, J.M. (1996). Comparaison de quelques méthodes de correction de la non-réponse d'un panel. *Techniques d'enquête*, 22, 43-53.

ROSENBAUM, P.R., et RUBIN, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70, 41-55.

SKINNER, C.J., et RAO, J.N.K. (1996). Estimation in dual frame surveys with complex designs. *Journal of the American Statistical Association*, 91, 433, 349-356.

STATISTIQUE CANADA (1995). Enquête nationale sur la santé de la population 1994-95 public use microdata file. No. 82F0001XCB au catalogue.

STUKEL, D.M., MOHL, C.A., et TAMBAY, J.L. (1997). Weighting for cycle two of Statistics Canada's National Population Health Survey. *Proceedings of the Survey Methods Section, Statistical Society of Canada*, 111-116.

TAMBAY, J.L., et CATLIN, G. (1995). Plan d'échantillonnage de l'Enquête nationale sur la santé de la population. *Rapports sur la santé*, Statistique Canada, 7, 31-42.

THOMSEN, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistik Tidskrift*, 11, 278-283.

ANGOSS SOFTWARE (1995). *Knowledge SEEKER IV for Windows - User's Guide*. ANGOSS Software International Limited.

CATLIN, G., et WILL, P. (1992). Enquête nationale sur la santé de la population: premiers faits saillants. *Rapports sur la santé*, Statistique Canada, 4, 313-319.

CZAJKA, J.L., HIRABAYASHI, S.M., LITTLE, R.J.A., et RUBIN, B.R. (1992). Projecting from advance data using propensity modelling: an application to income and tax statistics. *Journal of Business & Economic Statistics*, 10, 117-131.

DAVID, M.H., LITTLE, R., SAMUEL, M., et TRIEST, R. (1983). Imputation models based on the propensity to respond. *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, 168-173.

DUFOR, J., GAGNON, F., MORIN, Y., RENAUD, M., et SÄRNDAAL, C.-E. (1998). Measuring the impact of alternative weighting schemes for longitudinal data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*. À paraître.

ERNST, L. (1989). Weighting issues for household and family estimates. Dans *Panel Surveys*, (Éds. D. Kasprzyk, G. Duncan, G. Kalton et M.P. Singh). New York: John Wiley and Sons, 135-159.

GRONDIN, C. (1996). Pondération longitudinale – Première vague de l'EDTR. Note interne, division des méthodes d'enquête sociales, Statistique Canada.

INSTITUTE FOR SOCIAL RESEARCH (1979). A Panel Study of Income Dynamics: Procedures and Tape Codes – 1978 Interviewing Year – Wave XI – A Supplement. The University of Michigan.

KALTON, G., et KASPRZYK, D. (1986). Le traitement des données d'enquête manquantes. *Techniques d'enquête*, 12, 1-17.

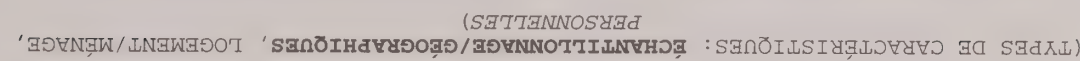


Figure 3. Classes de réponse à l'échelle provinciale, établies pour la non-réponse au cycle 2

Par ailleurs, le calendrier très serré auquel nous étions soumis nous a obligés à laisser de côté certaines analyses. Il aurait été intéressant d'approfondir les possibilités offertes par l'algorithme CHAD. Par exemple, comme ce dernier permet d'utiliser une variable de réponse nominale, nous aurions pu classer les unités d'échantillonnage en trois groupes, soit les déclarants vivants, les déclarants décédés ou hors du champ de l'enquête, et les non-déclarants. Nous aurions aussi aimé comparer une analyse effectuée avec l'algorithme CHAD et une autre avec la méthode de la régression logistique. Nous aurions pu utiliser les variables du volet Santé, telles que l'indice de santé ou le statut fumeur/consommateur d'alcool, pour définir les cellules d'ajustement de la pondération, bien que leur utilité aurait été amoindrie du fait de leur absence dans certaines unités (elles sont absentes des schémas de réponse G*-GS, G*-G* et G*-**). Il faudrait reconsidérer la décision d'utiliser les mêmes cellules d'ajustement de la pondération pour différents types de non-réponse. Aurait-il mieux valu utiliser les cellules d'ajustement créées pour la réponse au niveau des ménages pour l'analyse de la non-réponse transversale dans le volet Santé? Pour comparer l'efficacité des diverses stratégies d'ajustement de la non-réponse, il faudrait évaluer leur incidence sur la variance des estimateurs. Nous pourrions de la même façon évaluer l'incidence de l'ajustement de la non-réponse dans le cycle 2 sur le biais de la non-réponse, en utilisant les données du cycle 1 qui sont disponibles pour tous les panélistes. Les estimations portant sur l'ensemble de l'échantillon pourraient être comparées aux estimations ajustées en fonction de la non-réponse, générées à partir des unités déclarantes.

Le cycle 3 présentera de nouvelles difficultés. Un «aplanissement» de l'ensemble de l'échantillon est prévu pour cette année, qui influera sur la stratégie d'estimation transversale et, partant, sur le traitement de la non-réponse. Etant donné que la non-réponse dans le volet longitudinal augmentera, il faudra tenir compte des effets secondaires de l'ajustement de la pondération, tels que la création possible de valeurs aberrantes de pondération. Il deviendra ardu d'établir des groupes de pondération pour différents types d'analyses longitudinales, parce que le nombre de schémas de réponse partiellement augmentera. Combien de schémas peut-on traiter de front sans difficulté, et lesquels choisir? Il faudrait évaluer la pertinence des données supplémentaires, telles que le statut «personne déménagée», pour le traitement de la non-réponse. On devra probablement recourir à l'imputation pour la non-réponse dans le cycle 3. Comment pourra-t-on concilier le procédé d'imputation avec la méthode d'ajustement de la pondération dans le traitement de la non-réponse? Il reste beaucoup de travail autour de l'ENSP. Il faut espérer que nous aurons le temps de régler certaines questions en suspens avant que ne s'achève le traitement de la réponse pour le cycle 3.

4. CONCLUSION

a été sélectionnée au premier embranchement par l'algorithme CHAD. Cette constatation vient confirmer qu'il a été judicieux de prendre en compte les caractéristiques personnelles pour ajuster la non-réponse au niveau des ménages. Dans la figure 3b, Non-réponse longitudinale complète, la Région métropolitaine de recensement (CMA), l'état civil (MART) et le sexe, malgré leur importance moindre que les caractéristiques de revenu, de race et de lieu de naissance, ont été utilisées le plus souvent (cinq fois chacune). Comme il a déjà été souligné, les variables d'échantillonnage telles que les indicateurs filtrage (REJBCT) et ménages «Adulte/Enfant» (ACFLAG) ont été prises en compte afin que soit incorporé le plan d'enquête dans les analyses exécutées avec l'algorithme CHAD. Bien que ces variables aient été sélectionnées une fois seulement chacune, les caractéristiques du ménage utilisées dans le plan, telles que la présence d'enfants (KIDS) et de membres de moins de 25 ans (UND25) ont été sélectionnées à l'occasion. La taille du ménage n'a pas été utilisée, mais le type de famille (FAMTYP), une variable liée à la taille, a été sélectionné deux fois.

Les cellules d'ajustement produites par l'algorithme CHAD ont été révisées, mais très peu de modifications ont été apportées. À l'intérieur de chacune des cellules, le poids des unités déclarantes a été réparti au prorata, afin d'obtenir le poids total pour les unités de déclaration et de non-déclaration. L'ampleur de l'ajustement au poids des unités non déclarantes n'a jamais excédé 1,83.

Le présent document explique la stratégie utilisée pour le traitement de la non-réponse dans les volets longitudinal et transversal du cycle 2 de l'ENSP. L'approche adoptée permettait de tenir compte de considérations pratiques telles que la nécessité de se doter d'un moyen facile, valide sur le plan statistique, pour définir les cellules d'ajustement de la pondération et de produire un ensemble de données plus utile (par exemple, nous avons ajusté séparément la réponse longitudinale complète et partielle). Il fallait de plus éviter de compliquer inutilement le processus; c'est pourquoi nous avons utilisé à divers endroits les mêmes cellules d'ajustement de la pondération. Nous avons choisi l'algorithme CHAD au détriment de la régression logistique parce que la première approche nous donnait plus de liberté quant au nombre et à la diversité des variables pouvant être prises en compte. Ainsi, nous avons pu tenir compte de nombreuses variables d'échantillonnage et caractéristiques, pour en retenir quelques-unes seulement. Il nous semblait ainsi que, nous pourrions tirer profit de ces caractéristiques dans le traitement de la non-réponse.

CARACTÉRISTIQUES PERSONNELLES DU
PANÉLISTE

SEX	Sexe
AGE	Âge en années
AGE16	Indicateur personne âgée de 16 ans et plus
MARIT	Etat civil
FAMID	Identificateur famille à l'intérieur du ménage (A, B, C, ...)
RACE	Blanc, Noir, Autochtone ou autre
BORN	Lieu de naissance (Canada, E.-U./Mexique
AGIMM	Amérique S./Afrique, Europe/Australie, Asie)
	Âge au moment de l'immigration pour les immigrants
*MOVED	Indicateur changement de province (voir le texte)
*EDUC	Niveau de scolarité le plus élevé (12 catégories)

*STUDENT	Indicateur étudiant (12 catégories)
MACT	Activité principale (8 catégories)
*NUMJOB	Nombre d'emplois occupés l'an dernier (cycle 1)
RESTR	Indicateur pour activité restreinte
*CAUSE	Cause principale de restriction (12 catégories)
CONSUL	Nombre de consultations auprès d'un médecin
INHOSP	Indicateur pour séjour de 1 nuit à l'hôpital
*CHRONIC	Nombre de maladies chroniques

* Indique que la variable n'a jamais été significative lors de l'établissement de classes.

La figure 3 énonce les variables choisies par l'algorithme CHAID en vue de créer des cellules d'ajustement de la non-réponse pour la réponse au niveau des ménages et pour la réponse longitudinale complète dans chaque province. À des fins de confidentialité, on ne peut donner le détail sur les tailles des cellules individuelles et les taux de réponse (certaines des variables utilisées sont considérées comme étant sensibles et ne figurent pas dans les FMGD). Cependant, on trouvera de l'information sommaire sur la construction des cellules aux tableaux 2 et 3.

Les résultats varient d'une province à l'autre. Comme il était prévu, les provinces où l'échantillon est plus britannique et l'Alberta, ont permis de tracer des arbres de décision «plus riches». Les tailles des cellules ainsi que les taux de réponse varient aussi passablement. Dans le tableau 2, qui porte sur la réponse au niveau des ménages, une seule cellule est attribuée au Manitoba, et une seule cellule contient 88 % de l'échantillon du Nouveau-Brunswick. De même, dans le tableau 3, une grande partie de l'échantillon de Terre-Neuve est placée dans une seule

des deux cellules. On constate pour quelques provinces des cellules avec des taux de non-réponse autour de 40 %.

La figure 3 montre les différences entre les caractéristiques des classes de pondération pour les différentes provinces, mais aussi pour les deux types de non-réponse à l'intérieur des provinces. Dans toutes les provinces, sauf en Alberta, l'algorithme CHAID utilise des caractéristiques distinctes pour les deux types de non-réponse, dès le premier ou le deuxième niveau d'embranchement. Quelques caractéristiques sont dominantes aux premiers niveaux d'embranchement dans de nombreux arbres, pour les deux types de non-réponse, soit le niveau de suffisance du revenu du ménage (INCNR), l'indicateur de non-réponse quant au revenu (INCNR), la race (RACE), et le lieu de naissance (BORN).

Caractéristiques des cellules d'ajustement de la réponse (réponse au niveau des ménages)											
Prov.	Nbre	Nbre	Tailles cellules	% par cellules NR	min.	max.	moy.	min.	max.	moy.	
T.-N.	1 082	40	354	728	541	1,4	4,8	3,7	4,9	10,9	5,1
I.P.-E.	1 037	51	81	478	259	3,0	13,6	4,9	5,1	10,9	5,1
N.-E.	1 085	55	46	374	217	0,7	10,9	5,1	5,2	34,4	5,2
N.-B.	1 125	59	32	986	281	2,6	34,4	5,2	4,4	12,1	4,4
QC	3 000	133	123	2 363	750	1,8	12,1	4,4	7,3	25,8	7,3
Ont.	4 307	315	44	1 038	308	0,9	25,8	7,3	4,1	4,1	4,1
Man.	1 205	50	1 205	1 205	1 205	4,1	4,1	4,1	5,1	35,3	5,1
Sask.	1 168	59	37	626	167	1,6	35,3	5,1	7,5	36,7	7,5
Alb.	1 544	116	32	837	221	3,9	29,0	8,6			
C.-B.	1 723	149	82	678	246	5,2	29,0	8,6			

Caractéristiques des cellules d'ajustement de la réponse (réponse longitudinale complète)											
Prov.	Nbre	Nbre	Tailles cellules	% par cellules NR	min.	max.	moy.	min.	max.	moy.	
T.-N.	1 082	73	35	1 047	541	6,2	22,9	6,7	7,7	26,8	7,7
I.P.-E.	1 037	80	41	453	207	4,1	26,8	7,7	8,8	14,3	8,8
N.-E.	1 085	96	236	555	362	6,5	16,8	7,6	7,0	37,8	7,0
N.-B.	1 125	86	59	819	281	4,8	16,8	7,6	10,9	38,0	10,9
QC	3 000	211	42	2 202	375	2,5	37,8	7,0	7,6	15,1	7,6
Ont.	4 307	470	34	619	196	0,0	38,0	10,9	7,1	28,9	7,1
Man.	1 205	91	186	763	402	5,6	15,1	7,6	9,6	39,0	9,6
Sask.	1 168	83	90	339	195	0,0	28,9	7,1			
Alb.	1 544	148	41	866	172	1,1	39,0	9,6			
C.-B.	1 723	192	33	408	191	4,5	37,3	11,1			

Dans la figure 3a, le revenu du ménage et les variables liées (INCNR et INC SRC), le statut propriétaire/locataire (OWN ER), la race, le lieu de naissance et le type de logement (DWELL) ont toutes été utilisées trois fois ou plus pour la création des classes de pondération de la non-réponse au niveau des ménages. On remarque aussi que dans cinq des neuf provinces, une caractéristique personnelle des panélistes

À la lumière des données du tableau, il faut construire des cellules d'ajustement pour cinq types de réponse (ou de non-réponse) dans le cycle 2, comme suit:

- FMGD Généralités – réponse des ménages
- FMGD Généralités – réponse des personnes
- FMGD Santé – réponse combinée
- FMGD longitudinal – réponse complète
- FMGD longitudinal – réponse partielle

Seulement trois groupes de cellules d'ajustement ont été créés pour ces types de réponse. Les cellules d'ajustement créées pour la réponse au niveau des ménages dans le FMGD Généralités ont aussi servi pour les réponses partielles dans le FMGD longitudinal, puisque la réponse partielle au niveau des ménages correspondait presque toujours à une réponse partielle au niveau des personnes dans le fichier longitudinal (sauf pour 53 individus). De même, les cellules d'ajustement créées pour la réponse complète dans le FMGD longitudinal ont été appliquées à la réponse dans le FMGD Santé. Bien qu'on ait obtenu 366 réponses de plus pour le dernier type (schéma G*-GS), on a considéré que le même mécanisme de réponse était à l'oeuvre. Le troisième groupe de cellules d'ajustement a été appliqué à la réponse au niveau des personnes dans le FMGD Généralités. On a utilisé les catégories province – âge – sexe, comme ce fut le cas au cycle 1.

Incidentement, bien que les mêmes cellules d'ajustement puissent être utilisées pour différents groupes de données, on effectue des calculs distincts pour chaque type de groupe de données. Ainsi, les 366 dossiers assortis d'un schéma de réponse G*-GS sont traités comme répondants dans le processus d'ajustement des poids pour le FMGD Santé, mais comme non-répondants dans le processus d'ajustement des poids pour l'analyse de la réponse complète dans le FMGD longitudinal.

3.4 Création de cellules d'ajustement de la pondération

On a créé des groupes distincts de cellules d'ajustement de la pondération pour chaque province. Tout d'abord, il a fallu déterminer les variables à prendre en compte. Avec l'approche CHAD, le nombre de variables à considérer importe peu, et on a tenu compte de différents types de variables. Bien entendu, on a pris en compte les caractéristiques du ménage, ou du logement, ainsi que les caractéristiques personnelles du panéliste. Afin d'être en mesure d'incorporer le plan d'enquête dans l'analyse, nous avons aussi tenu compte de certaines caractéristiques liées à ce plan ou au poids d'échantillonnage. Parmi celles-ci se trouvent les variables géographiques telles que le code régional métropolitain de recensement (RMR) ou l'indicateur urbain/rural, les variables d'échantillonnage particulières au cycle 1, telles que les indicateurs de ménages visés par le filtrage et le type de ménage «adulte/enfant», de même que les caractéristiques liées à

l'application de ces variables d'échantillonnage, telles que la présence dans un ménage d'un membre âgé de moins de 25 ans ou d'un enfant. La taille du ménage a été considérée comme étant une caractéristique du ménage, et on l'a aussi liée au poids de l'échantillon. À la lumière des expériences passées, on a décidé d'inclure, outre la caractéristique du revenu du ménage, une caractéristique fictive indiquant si le revenu du ménage avait été déclaré dans le cycle 1. Les changements d'adresse pouvant conduire à une incapacité de retracer la provenance d'une non-réponse, il aurait été utile d'intégrer un identificateur de changement d'adresse. Cependant, dans certains cas de non-réponse et d'absence de contact, il était difficile de déterminer si le panéliste avait vraiment déménagé. Enfin, nous avons utilisé pour l'analyse une variable «personne ayant déménagé», qui nous permet de savoir si le panéliste avait changé de province entre les cycles. Cette modalité représente un pis-aller parce que la valeur par défaut est «non». Des caractéristiques personnelles tirées du volet Santé de l'enquête, telles que le statut fumeur/consommateur d'alcool, le niveau d'indice de santé et l'échelle de santé/souffrance morale n'ont pas été utilisées parce qu'elles étaient inconnues pour quelque 500 panélistes.

On trouve ci-après la liste des variables qui ont été utilisées. La valeur de l'indicateur de non-réponse, soit la variable dépendante, a été assignée en fonction de la définition donnée au concept de non-réponse.

VARIABLES D'ÉCHANTILLONNAGE/GÉOGRAPHIQUES

PROVINCE	Analyse effectuée à l'échelle provinciale
CMA	Région métropolitaine de recensement (0 s'il ne s'agit pas d'une RMR)
URBAN	Indicateur urbain/rural
REJECT	Indicateur qu'une unité (ménage) est sujette au filtrage
ACFLAG	Classe d'échantillonnage Adulte/enfant pour l'unité

CARACTÉRISTIQUES DU LOGEMENT/MÉNAGE

DWELL	Type de logement (10 catégories)
OWNER	Indicateur propriétaire/locataire
FAMTYP	Type de famille (personne seule, famille monoparentale, couple marié, autre)
INC	Suffisance du revenu du ménage (5 niveaux)
INCNR	Indicateur de non-réponse pour le revenu
INCSRC	Principale source de revenu (6 catégories)
*HHSIZE	Taille du ménage
UND25	Indicateur membre de moins de 25 ans
KIDS	Indicateur enfant de moins de 12 ans

CHAD ne les retient tout simplement pas. Pour le volet Généralités, la non-réponse au niveau individuel indique qu'une information est disponible pour certains membres seulement d'un ménage, en raison d'un refus de l'un des membres ou d'une absence temporaire. Le taux de non-réponse au niveau individuel étant de 1,8 % seulement, il a été décidé que la création de cellules de pondération en fonction des catégories province, âge et sexe (comme dans le Cycle 1) suffirait amplement.

Contrairement au FMGD Généralités, on a pu combiner l'ajustement de la non-réponse aux niveaux des ménages et des personnes pour les FMGD longitudinal et Santé, parce qu'il est appliqué aux réponses d'une même personne (le paneliste). On a donc créé un seul ensemble de cellules d'ajustement.

Pour ce qui est du FMGD longitudinal, on a noté que les éléments de données provenaient à la fois des volets Généralités et Santé, mais que les taux de réponse différaient d'un volet à l'autre. Cet écart a entraîné des schémas de réponse différents pour les cycles 1 et 2, comme suit: GS-GS; GS-G*; GS-GS; GS-G*. À ces différences s'ajoutent les schémas de non-réponse longitudinale GS-G* et GS-G** (les lettres correspondent aux réponses pour les volets dans chaque cycle, et l'astérisque correspond à une non-réponse). Afin de maximiser l'utilité des données, on a décidé d'ajuster de deux façons la non-réponse longitudinale. L'une des méthodes d'ajustement est appliquée au schéma «réponse longitudinale complète» GS-GS. Autre-ment dit, tous les autres schémas de réponse sont considérés comme étant des non-réponses. L'autre méthode d'ajustement s'applique à la «réponse longitudinale partielle», pour les cas où on a obtenu au moins les données générales pour chacun des cycles (schémas GS-GS; GS-G*; GS-GS et GS-G*). Le groupe de données Réponse complète sera utile aux chercheurs qui souhaitent analyser un ensemble de données longitudinales complètes couvrant tout le contenu du questionnaire. Le groupe de données Réponse partielle peut servir aux chercheurs qui s'intéressent surtout aux types de variables du questionnaire général. On le constate dans le tableau ci-dessous, la Réponse longitudinale partielle dépasse de 3 % seulement la Réponse longitudinale complète.

Tableau 1

Schémas de la réponse longitudinale			
Type de réponse	Schéma de	Nombre	
Cycles 1-2			
de dossiers			
Complète			
Partielle			
	GS-GS	15 670	■
	GS-G*	110	■
	GS-GS	366	■
	G*-GS	22	■
	GS-G*	1 014	■
	GS-G**	94	■
	G*-**	17 276	■
Total			

3.3 Ajustement de la non-réponse dans l'échantillon principal

Il a fallu établir une méthode d'ajustement de la non-réponse pour chacun des FMGD, soit les fichiers de données longitudinales, Généralités (transversales) et Santé (transversales). Le FMGD Généralités sera traité en premier.

Comme le démontre la figure 2, la stratégie de pondération appliquée au FMGD Généralités a exigé d'ajuster séparément la non-réponse aux niveaux des ménages et des personnes. Quand on a créé des cellules d'ajustement de la non-réponse au niveau des ménages, on a considéré les caractéristiques des panelistes ainsi que celles du ménage comme étant des prédicteurs de non-réponse, pour trois raisons. Premièrement, étant donné que le paneliste représentait le lien avec le ménage dans le cycle 2, on considèrerait que ses caractéristiques avaient permis de trouver le ménage et d'obtenir une réponse (cette personne a souvent permis d'établir le premier contact). Deuxièmement, quelques caractéristiques personnelles du paneliste, telles que la race, constituent jusqu'à un certain point des caractéristiques du ménage. Enfin, l'utilisation des caractéristiques du paneliste n'était pas incompatible avec la nécessité qui nous incombat d'utiliser divers renseignements pour construire les cellules de pondération. Si les caractéristiques

programme, il vaut mieux limiter le nombre de variables et de termes d'interaction. La fusion des cellules peut aussi s'avérer complexe, comme ce fut le cas pour l'EDTR, déjà mentionnée. Au contraire, l'algorithme CHAD permet de traiter un grand nombre de covariables et, en raison de la structure en arbre de décision, il est facile d'établir les interactions. Qui plus est, il est facile d'incorporer les exigences quant à la taille minimale des cellules comme paramètres d'exécution de programme. Parmi les principaux inconvénients, citons le fondement théorique moins familier (le logiciel Knowledge Seeker est vendu comme un outil d'«exploration des données»), la documentation limitée et le peu de logiciels d'exécution. À ceux-ci s'ajoute l'impossibilité d'incorporer le plan d'échantillonnage pour adapter les modèles logistiques aux données de l'enquête, comme c'est le cas pour certains programmes statistiques tels que SUDAAN et PC CARP. Pour pallier à ces contraintes dans l'ENSP, on a inclus comme prédicteurs des caractéristiques liées au plan d'échantillonnage (se reporter à la section 3.4).

Deux études empiriques comparant les approches logistiqu et CHAD pour l'analyse de la non-réponse ont obtenu des résultats divergents. Ainsi, Rizzo, Kalton et Brick (1996) n'ont pas conclu à des différences notables entre les deux approches appliquées à la Survey of Income and Program Participation. Par contre, Dufour, Gagnon, Morin, Renaud et Sarnadal (1998), dans une étude simulée de l'EDTR, ont trouvé que l'application de l'approche CHAD pour l'ajustement de la non-réponse produisait un biais moins important.

admissibles

transversale.

Aux fins de l'étude transversale, tous les cas présentés au paragraphe précédent ont été traités comme des situations ne faisant pas partie du champ de l'enquête. Ce choix est acceptable parce que des outils d'enquête distincts pour les établissements et pour les territoires ont été utilisés pour la couverture transversale de ces populations en particulier, qu'il s'agisse du champ de l'enquête. Les unités non admissibles ne font pas partie de l'ENSP, mais, comme elles représentent d'autres unités de ce genre, elles ont été traitées aux fins de la pondération comme des répondants à toutes les étapes de l'ajustement de la pondération, sauf celle de l'intégration et celle de la stratification a posteriori.

Les refus et les cas où des questionnaires étaient manquants pour des raisons autres que celles données précédemment ont été définis comme des non-réponses. Comme nous le verrons, une distinction a été faite ultérieurement entre les non-répondants complets ou partiels afin de tenir compte de différents utilisateurs.

pour l'ajustement de la non-réponse

de tenir compte de différents utilisateurs.

Dans l'approche de modélisation des segments, on génère un arbre de décision à partir des données; pour ce faire, on divise les ensembles de données en séquence de façon que, à chaque noeud, le prédicteur le plus significatif pour la variable de la réponse serve de point de départ à la division suivante. La division se poursuit jusqu'à ce qu'il ne soit plus possible de trouver une variable significative pour effectuer une autre division, ou qu'il ne soit plus possible de respecter les exigences quant à la taille minimale de la cellule. L'une des premières applications de la méthode de modélisation des segments pour l'ajustement de la non-réponse portait sur la Panel Study of Income Dynamics (Institute for Social Research 1979). En raison de ses avantages, énoncés ci-après, on a adopté pour l'ENSP l'approche de modélisation des segments fondée sur l'algorithme CHAD (détection automatique d'interactions du chi carré), inventé par Kass (1980). L'algorithme CHAD utilise des tests du χ^2 pour diviser les prédicteurs catégoriques et retenir la division la plus significative à chaque étape. La division, en deux catégories ou plus, n'est pas exécutée de la même façon pour les prédicteurs ordonnés et non ordonnés. L'algorithme CHAD a été appliqué à l'aide du logiciel Knowledge Seeker (ANGOSS Software, 1995). Il faut noter que le logiciel applique l'algorithme aux prédicteurs continus en les traduisant tout d'abord en

ար ուսուցիչը պետք է խոստովանի իր սխալները և խնայողությամբ օգտագործի իր ժամանակը:

2.4 Données de sortie et pondération du cycle 2

La figure 2 constitue un sommaire des trois principaux fichiers de sortie prévus pour le cycle 2: un FMGD pour l'étude longitudinale; un FMGD pour le volet Santé de l'étude transversale. Le FMGD de l'étude longitudinale contient des données sur les volets Généralités et Santé pour les deux cycles pour les 17 000 panelistes [nota: pour des raisons de confidentialité, il est possible que nous ne puissions publier le FMGD de l'étude longitudinale – auquel cas seul un fichier interne de microdonnées serait produit]. Le FMGD du volet Santé de l'étude transversale contient les renseignements de 1996 sur les volets Généralités et Santé pour environ 70 000 panelistes et membres sélectionnés par GANT. Le FMGD du volet Généralités de l'étude transversale contient les renseignements de 1996 sur le volet Généralités pour environ 220 000 membres de l'échantillon principal et du groupe GANT. Les procédés de pondération propres à chaque FMGD, présentés ci-après pour l'échantillon principal, sont décrits de façon plus détaillée dans l'étude de Stukel, Mohl et Tambay (1997).

Fichier de sortie	Contenu	Échantillons	Unités	Taille	Pondération (échantillon principale)	1. Pondération pour non répondants	2. Ajust. panelistes non répondants	3. Intégration échant. principal et GANT	4. Ajust. membres des ménages NR	5. Intégration échant. principal et GANT	6. Stratification a postériori
FMGD	Généralités et Santé	Principal seulement	Paneliste	≈ 17 000 dossiers	1. Pondération pour non répondants	2. Ajust. panelistes non répondants	3. Intégration échant. principal et GANT	4. Ajust. membres des ménages NR	5. Intégration échant. principal et GANT	6. Stratification a postériori	
FMGD – ÉTUDE LONGITUDINALE	Généralités et Santé	Principal et GANT (3 prov.)	Panelistes et GANT	≈ 70 000 dossiers	1. Pondération pour l'année de base	2. Ajust. ménage non répondants	3. Ajust. répartition du poids	4. Ajust. membres des ménages NR	5. Intégration échant. principal et GANT	6. Stratification a postériori	
FMGD – ÉTUDE TRANSVERSALE	Généralités seulement	Principal et GANT (3 prov.)	Tous les membres des ménages	≈ 220 000 dossiers	1. Pondération pour l'année de base	2. Ajust. ménage non répondants	3. Ajust. répartition du poids	4. Ajust. membres des ménages NR	5. Intégration échant. principal et GANT	6. Stratification a postériori	

Figure 2. Description des fichiers de sortie pour le cycle 2

On obtient les pondérations pour le FMGD de l'étude longitudinale en ajustant la pondération pour l'année de base d'abord pour les non-réponses du groupe de 1996, puis pour la stratification a postériori. La pondération pour l'année de base représente la fraction de sondage inverse pour 1994, y compris tous les ajustements pour le FMGD du volet Santé décrits à la section 2.2 jusqu'à l'ajustement (4) pour la sélection du panel (une correction est nécessaire pour le «retrait» des additions provinciales à l'échantillon de 1994). Le facteur de compensation de la non-réponse est traité dans la prochaine section et y sera décrit. La stratification a postériori est effectuée pour reproduire les chiffres de 1994 pour la population des provinces selon l'âge et le sexe.

Pour ce qui est du FMGD du volet Santé de l'étude transversale, le processus de pondération pour les panelistes (échantillon principal) nécessite trois ou quatre étapes. Habituellement, la pondération de l'année de base est ajustée pour la non-réponse des panelistes, selon les explications qui se trouvent dans la prochaine section, ainsi que pour la stratification a postériori (pour établir la correspondance avec les chiffres de population régionaux ou provinciaux de 1996, selon l'âge et le sexe). Dans les provinces où une partie de l'échantillon provient de la GANT, l'étape supplémentaire consiste à intégrer l'échantillon composé de répondants sélectionnés par la GANT. On obtient l'estimation intégrée par une forme d'adaptation minimale de l'estimateur double Skinner-Rao (Skinner Rao 1996). Quant au FMGD du volet Généralités de l'étude transversale, le processus de pondération pour l'échantillon principal nécessite cinq ou six étapes. Premièrement, une fois de plus, il faut calculer la pondération de l'année de base. Il faut ensuite procéder à un ajustement pour la non-réponse au niveau des ménages, dont traitera la prochaine section. La prochaine étape consiste à appliquer la méthode de répartition du poids. Cette méthode a été décrite par Ernst (1989) et mise au point plus tard par Lavallée (1995). Le poids du paneliste divisé par le nombre de personnes de son ménage qui étaient admissibles au cycle 1 est attribué à tous les membres du ménage, y compris ceux qui n'étaient pas admissibles au cycle 1 (c'est-à-dire nouveaux-nés, immigrants). Cette méthode est non biaisée pour les estimations des totaux pour la population des ménages où au moins un membre était admissible au cycle 1. La prochaine étape consiste à établir l'ajustement pour la non-réponse d'un membre du ménage. Dans les provinces qui ont eu recours à la GANT, cette étape est suivie de l'intégration de l'échantillon principal avec l'échantillon GANT (cette fois pour tous les âges). La stratification a postériori est effectuée de la même façon que pour le FMGD du volet Santé de l'étude transversale.

3. STRATÉGIE RELATIVE À LA NON-RÉPONSE DANS L'ÉCHANTILLON PRINCIPAL DU CYCLE 2

La présente section explique la stratégie adoptée pour le traitement de la non-réponse du cycle 2 de l'échantillon principal (non GANT). L'ajustement pour la non-réponse a encore une fois été effectué à l'aide de la méthode de la cellule de pondération sauf que, cette fois-ci, les données du cycle 1 étaient disponibles, ce qui a permis de créer des cellules plus homogènes en regard de la propension à répondre. Dans la section 3.1, on définit les non-répondants de l'ENSP. Dans la section 3.2, on explique de deux méthodes générales utilisées pour créer des cellules de pondération, dont celle choisie pour l'ENSP. La stratégie d'ajustement pour la non-réponse est expliquée dans la section 3.3 alors qu'à la section 3.4, on décrit la création des cellules de pondération de la non-réponse.

des personnes [au sein des ménages répondants]; 4) un ajustement simple pour la stratification a posteriori. L'ajustement 2) n'a été appliqué qu'aux ménages qui n'ont aucun membre de moins de 25 ans. La formule utilisée est $1/(1 - r_s)$, où r_s représente la fraction de sondage du sous-échantillon pour le filtrage appliqué dans la strate. L'ajustement pour la stratification a posteriori a été effectué séparément pour chaque classification croisée province-âge groupe-sexe. Les pondérations résultant de toutes les étapes précédentes sont multipliées par le rapport entre la population connue et la population estimée au sein de la classification croisée. Les chiffres relatifs à la population connue sont en fait des projections découlant du recensement.

Les ajustements apportés relativement à la non-réponse au niveau des ménages et des personnes (1) 1,3 % et 1,4 %, respectivement) ont été appliqués aux unités déclarantes puisque les non-répondants ont été exclus des FMGD. Si w_i est le poids de l'échantillon d'une unité i , le poids ajusté de la non-réponse, $w_{aj,i}$ se définit comme $w_{aj,i} = w_i (\sum_{\text{tout}} w_j) / (\sum_{\text{rép}} w_j)$, où les sommes s'effectuent sur toutes les unités d'échantillonnage et toutes les unités déclarantes, respectivement, à l'intérieur des cellules de pondération pour l'ajustement de la non-réponse. En raison d'un manque de renseignements de la part des ménages non répondants, les cellules de pondération pour la non-réponse des ménages ne sont que de simples classifications croisées des strates de l'ENSP et de la saison (c'est-à-dire, les groupes d'être par opposition aux groupes d'hiver). Les cellules de pondération pour la non-réponse des individus, dont le taux est très faible, sont les classifications croisées province-âge-sexe qui ont été utilisées pour l'ajustement de la stratification a posteriori.

Les ajustements apportés aux poids pour le FMGD du volet Santé comprennent: 1) un ajustement pour la non-réponse des ménages; 2) un ajustement pour la procédure de filtrage; 3) un ajustement pour le sous-échantillon des ménages «enfants» et des ménages «adultes»; 4) un ajustement pour la sélection des panélistes de l'enquête longitudinale; 5) un ajustement pour la non-réponse des panélistes (6) un ajustement pour la stratification a posteriori. Les deux premiers ajustements sont exactement les mêmes que pour le FMGD du volet Généralités. Comme le FMGD du volet Santé n'inclut pas les membres du groupe qui sont des enfants, l'ajustement 3) compense pour les ménages de l'échantillon où des membres non-enfants ne sont pas admissibles comme panélistes. Conséquemment, l'ajustement ne s'applique qu'aux ménages avec enfants et équivalait à $1/r$, où r représente la proportion de ménages «adultes» dans l'échantillon. L'ajustement 4) est l'inverse de la probabilité que l'on ait sélectionné le panéliste. Les ajustements pour la non-réponse des panélistes (3,9 %) et pour la stratification a posteriori sont similaires à ceux des FMGD pour le volet Généralités, et utilisent les mêmes classifications croisées province-âge-sexe. Même si les panélistes enfants ne sont pas inclus dans les FMGD du volet Santé, aux fins de l'enquête longitudinale, leur

immigrants.

Nous faisons remarquer, qu'aux fins de l'étude transversale, l'échantillon principal ne couvre pas très bien les nouveaux arrivants dans la population comme les nouveaux-nés et les immigrants récents. La population à qui on a administré le volet Généralités du questionnaire est constituée des résidents des ménages où au moins un des membres était admissible au cycle 1; les ménages composés entièrement de récents immigrants (et de leurs nouveaux-nés) sont donc exclus. La population à qui on a administré le volet Santé du questionnaire est constituée de personnes qui étaient admissibles au cycle 1: les immigrants récents et les enfants de moins de deux ans sont exclus de la population cible de l'échantillon principal (ils sont inclus dans la population cible pour la GANT). Pour les deux questionnaires, l'ajustement effectué en utilisant des chiffres sur la population qui n'excluent pas les récents immigrants. Il en résulte que la population de récents immigrants est implicitement estimée par la population de non-immigrants parce que les dernières pondérations de l'échantillon principal sont ajustées à la hausse pour rendre compte des chiffres précédents. Il s'agit d'une contrainte qui est indiquée dans la documentation des FMGD. D'autres méthodes auraient pu être envisagées comme la stratification a posteriori en utilisant uniquement la population non immigrante ou en ajustant d'une façon quelconque seulement les pondérations des immigrants (qui sont récents) pour tenir compte des immigrants plus récents (qui ne le sont pas). Ces méthodes auraient été difficiles à appliquer lorsque, dans le cadre du questionnaire Généralités, une distinction doit être établie entre les ménages composés uniquement d'immigrants et les ménages mixtes dont certains membres sont des immigrants.

2.3 Plan d'échantillonnage du cycle 2

Pour le cycle 2, l'accent a davantage été mis sur l'estimation longitudinale: aucune «bonification» de l'échantillon n'a été prévue avant le cycle suivant. L'échantillon principal était donc constitué d'environ 17 000 panélistes et de leur ménage actuel. Les panélistes ont été suivis et on leur a administré les volets Généralités et Santé du questionnaire, alors que les autres membres du ménage n'ont répondu qu'au volet Généralités. Aucun suivi n'a été effectué auprès des ménages non répondants de 1994. En Alberta, au Manitoba et en Ontario d'importants échantillons supplémentaires (ne faisant pas partie de l'échantillon principal) ont été obtenus par génération aléatoire de numéros de téléphone (GANT), ce qui a permis la production d'estimations transversales au niveau infraprovincial. Pour chaque ménage contacté par GANT, un membre de plus de 12 ans a été choisi pour remplir le volet Santé du questionnaire. En Alberta et au Manitoba, dans les ménages contactés par GANT qui avaient des enfants, un enfant a aussi été choisi pour remplir le volet Santé.

2. VUE D'ENSEMBLE DU PLAN D'ÉCHANTILLONNAGE DE L'ENSP ET DES DONNÉES DE SORTIE

2.1 Plan d'échantillonnage du cycle 1

Le premier échantillon de ménages a été constitué en 1994 à l'aide de l'outil de sélection d'un échantillon élaboré pour l'Enquête sur la population active (EPA) et, dans la province de Québec, en utilisant les logements qui ont participé à une enquête sur la santé menée par Santé Québec l'année précédente. Dans les deux cas, les ménages ou les logements ont été choisis au hasard dans des échantillons stratifiés de groupes sélectionnés selon la méthode de la probabilité proportionnelle à la taille. Les groupes ont été organisés en échantillons répétés et d'une période de collecte pour tenir compte du caractère saisonnier et pour permettre l'estimation de la variance. Il y a eu deux périodes de collecte estivale (juin et août) et deux périodes de collecte hivernale (novembre et mars 1995).

La figure 1 illustre le mécanisme de sélection du panel appliqué à l'extérieur du Québec. Les ménages de l'échantillon ont été désignés aléatoirement comme des ménages «adultes» ou «enfants», et comme admissibles ou non au filtrage, avant la collecte. Le filtrage a provoqué une augmentation dans le panel du nombre d'habitants provenant de ménages plus nombreux qui seraient sous-représentés si l'on choisissait seulement un membre par ménage, particulièrement les enfants et les adolescents. Les ménages admissibles au filtrage ont été rejetés de l'échantillon si aucun de leur membre n'avait moins de 25 ans. Le filtrage n'a pas été utilisé au Québec puisque les données provenant de l'enquête provinciale sur la santé permettaient l'application de différents taux de sous-échantillonnage par type de ménage et par taille.

Type d'unité de sondage	Caractéristique de sélection du panel	Limitée à:
Ménages	Aucun membre < 25 ans	s.o. – ménage rejeté
«enfants» admissibles au filtrage (AF)	Pas d'enfant, quelques membres < 25 ans	N'importe quel membre
Ménages «enfants» non AF	Présence d'enfants	Enfants
Ménages «adultes»	Tous	Membres de plus de 12 ans

Figure 1. Mécanisme de sélection du panel à l'extérieur du Québec

La classification en ménages «adultes» et «enfants» a été effectuée pour des raisons opérationnelles: le questionnaire sur la santé, pour les enfants, ne serait pas disponible avant

les périodes de collecte hivernales. Pour ce qui est des ménages «adultes», qui pouvaient être interviewés en tout temps, les enfants de moins de 12 ans n'étaient pas admissibles dans ce groupe. Les ménages «enfants», même ceux des groupes de la période estivale, ont été interviewés lors d'une période de collecte hivernale. Si des enfants étaient présents dans ces ménages, alors la sélection était limitée à eux. Pour réduire les distorsions saisonnières sur la charge de travail relative à la collecte de données et sur la représentativité du groupe découlant de ces procédures, un nombre moins grand de ménages ont été classés ménages «enfants» dans les groupes de la période estivale et, sauf une exception mineure, le filtrage a été appliqué seulement aux ménages «enfants».

Les provinces souhaitant améliorer les estimations au niveau infraprovincial pouvaient financer l'ajout d'effectifs à l'échantillon. Trois provinces se sont prévaluées de cette possibilité en augmentant la taille de l'échantillon dans des régions ciblées. En Colombie-Britannique, un échantillon additionnel de 800 ménages a été constitué dans un établissement local de santé en utilisant le système de génération aléatoire de numéros de téléphone (GANT). La taille prévue de l'échantillon total dans les provinces était d'environ 23 000 ménages après le filtrage.

Ce qui précède donne une indication générale du plan d'échantillonnage de 1994, et ces renseignements sont suffisants pour les besoins du présent document. Les lecteurs désireux d'obtenir une présentation plus précise de l'échantillon de 1994 peuvent consulter Tamby et Catlin (1995) ou Statistique Canada (1995).

2.2 Pondération et résultats pour le cycle 1

Le principal résultat de l'ENSP consiste en un ensemble de fichiers de microdonnées à grande diffusion (FMGD) des réponses individuelles confidentielles (des versions internes de ces fichiers dont certains renseignements sont supprimés pour assurer la confidentialité sont aussi créés). Pour 1994, un FMGD pour le volet Généralités (58 400 dossiers) et un FMGD pour le volet Santé (17 600 dossiers) contenant les données recueillies auprès de chacun des membres des ménages et des membres sélectionnés dans le groupe sans enfants respectivement ont été publiés (Statistique Canada 1995).

Les poids propres à chaque dossier contenu dans un FMGD ont été calculés en appliquant une série d'ajustements à un poids de base représentant les fractions de sondage inverse (FSI) des ménages. Les FSI sont calculées en multipliant les poids des FSI originales ou des échantillons de Santé Québec par l'inverse de la fraction de sondage du sous-échantillon utilisé par l'ENSP. Pour ne pas alourdir le texte, nous décrivons seulement les principaux ajustements utilisés à l'extérieur du Québec.

Les ajustements apportés aux poids pour le FMGD du volet Généralités comprennent: 1) un ajustement pour la non-réponse des ménages; 2) un ajustement pour la procédure de filtrage; 3) un ajustement pour la non-réponse

Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population

JEAN-LOUIS TAMBAY, IOANA SCHIOPU-KRATINA, JACQUELINE MAYDA,
DIANA STUKEL et SYLVAIN NADON¹

RÉSUMÉ

L'Enquête nationale sur la santé de la population (ENSP) est l'une des trois principales enquêtes-ménages longitudinales que mène Statistique Canada à une grande échelle auprès de la population canadienne. Depuis vingt ans, tous les deux ans, on a suivi un panel constitué d'environ 17 000 personnes. Les données provenant de l'enquête sont utilisées pour des analyses longitudinales, même si l'un des objectifs importants est la production d'estimations transversales. Pour chaque cycle, les panelistes fournissent des renseignements détaillés sur leur santé (S) pendant qu'au même moment, pour augmenter l'échantillon transversal, des données socio-démographiques et quelques renseignements sur la santé sont recueillis (G) auprès de tous les membres des ménages. Cette stratégie de collecte présente différents schémas de réponse pour les panelistes après deux cycles: GS-GS, GS-G*, GS-**, G*-GS, G*-G*, où «*» indique une portion de données manquantes. Le présent article explique la méthodologie élaborée pour traiter ces types de non-réponse longitudinale de même que la non-réponse d'une perspective transversale. L'utilisation de facteurs de pondération pour la non-réponse et la création de cellules d'ajustement pour la pondération à l'aide de l'algorithme CHAID sont expliquées ici.

MOTS CLÉS: Enquêtes longitudinales; traitement de la non-réponse; algorithmes CHAID.

1. INTRODUCTION

En 1996-1997, Statistique Canada a terminé la collecte de données pour le cycle 2 de l'Enquête nationale sur la santé de la population (ENSP). Cette enquête longitudinale a été lancée en 1994 afin d'obtenir des renseignements complets sur l'état de santé de la population canadienne ainsi que sur les facteurs déterminants en matière de santé. La population de l'enquête comprend des résidents de ménages et d'établissements de santé de l'ensemble du pays. Dans les provinces, le questionnaire rempli par les ménages se compose de deux principaux volets qui sont administrés à l'aide d'interviews assistées par ordinateur. Le volet Généralités sert à recueillir des données socio-démographiques et des renseignements de base sur la santé pour chaque membre du ménage. Le volet Santé sert à obtenir des données plus détaillées sur la santé du membre du ménage choisi pour participer au groupe longitudinal. Même si l'ENSP est une enquête longitudinale, ses objectifs incluent la production d'estimations transversales périodiques (Catlin et Will 1992). La méthodologie de collecte des données reflète les besoins à la fois sur les plans longitudinal et transversal. Les panelistes choisissent pour le cycle 1, sont suivis tous les deux ans pendant une période de 20 ans. Les personnes résidant avec les panelistes durant ces périodes fournissent des données pour le volet Généralités qui sont utilisées pour l'estimation transversale. Le champ d'observation de la composante transversale de l'enquête se rétrécissant avec le temps, il faut procéder régulièrement à une «bonification» de l'échantillon. La première bonification est prévue pour le cycle 3, en 1998.

Le présent document explique la méthodologie utilisée dans le Cycle 2 pour traiter la non-réponse au niveau des ménages et des personnes (avec indicateur pour la non-réponse à une question). La méthodologie se fonde sur la pondération des répondants au sein des sous-populations appelées cellules de pondération afin de tenir compte de la non-réponse. La pondération est une méthode couramment utilisée pour le traitement de la non-réponse à une question. Le biais et la variance découlant de cette méthode ont été étudiés par Thomsen (1973), Oh et Scheuren (1983), Kalton et Kasprzyk (1986) et Little (1986), entre autres. Si les cellules de pondération sont définies de telle sorte que la non-réponse survient presque totalement de façon aléatoire au sein de chaque cellule, alors le biais dû à la non-réponse peut devenir négligeable. Dans la même veine, David, Little, Samuël et Tréist (1983) ont étendu à la non-réponse la théorie élaborée par Rosenbaum et Rubin (1983) dans le contexte du jumelage du degré de propension. Leurs résultats supposent que la pondération peut rajuster le biais de la non-réponse lorsque les cellules de pondération sont composées en fonction de la propension à répondre.

On trouve à la section 2, une vue d'ensemble du plan d'échantillonnage et des données de sortie pour les deux premiers cycles. La section 3 explique les stratégies de traitement de la non-réponse et leurs résultats. Quant à la section 4, on y trouve les remarques finales. Il faut noter que la méthodologie présentée se rapporte aux échantillons de ménages des provinces; cette méthodologie ne couvre par les échantillons des territoires et des établissements.

¹ Jean-Louis Tambay, Ioana Schiopu-Kratina, Jacqueline Mayda, Diana Stukel et Sylvain Nadon, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, 16^{ème} étage, Immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6.

et

$$\gamma_t^* = k_t^*(1 + \sum_{v=1}^{v=t+1} \prod_{n=v}^{n=t+1} (1 - k_n))$$

$$\gamma_{t'}^* = \sum_{v=t'+1}^{v=t'+1} \prod_{n=v}^{n=t'+1} (1 - k_n),$$

où

$$k_t = P_t^{\tau|t-1} / (P_t^{\tau|t-1} + 1),$$

et $P_t^{\tau|t-1}$, P_t^{τ} représentent les erreurs quadratiques moyennes de \hat{p}_t^{τ} , dans le cas de données jusqu'à $t - 1$, t respectivement, et ceux-ci sont estimés à l'aide du filtre de Kalman.

Ce résultat découle des équations qui sont utilisées pour

estimer μ_t^{τ} :

$$\hat{p}_t^{\tau} = k_t^{\tau} \bar{y}_t^{\tau} + (1 - k_t^{\tau}) \hat{p}_{t-1}^{\tau}$$

$$\hat{p}_{t-1}^{\tau} = k_{t-1}^{\tau} \bar{y}_{t-1}^{\tau} + (1 - k_{t-1}^{\tau}) \hat{p}_{t-2}^{\tau}$$

:

$$\hat{p}_{t'+1}^{\tau} = k_{t'+1}^{\tau} \bar{y}_{t'+1}^{\tau} + (1 - k_{t'+1}^{\tau}) \hat{p}_{t'}^{\tau}$$

(cf. Harvey 1990, équation 3.2.8), en exprimant chaque \hat{p}_t^{τ}

par $\bar{y}_t^{\tau}, \bar{y}_{t-1}^{\tau}, \dots, \bar{y}_{t'+1}^{\tau}, \hat{p}_{t'}^{\tau}$.

En comparant v et v , nous constatons de façon

empirique que

$$\sum_{i'}^i \gamma_2^* + \gamma_{i'}^* P_{i'-1}^{\tau-1} \approx t - t'.$$

Nous considérons ici le cas simple où $\text{var}(e_{it}^{\tau}) = \sigma_{e_{it}^{\tau}}^2$ et $\text{cov}(e_{it}^{\tau}, e_{it'}^{\tau}) = \rho \sigma_{e_{it}^{\tau}}^2$, avec $\rho \geq 0$, dans le cas de $t' \neq t$, c'est-à-dire où non seulement les variances, mais toutes les covariances aussi sont égales et non négatives. On peut alors montrer que

$$\sigma_{e_{it}^{\tau}}^2 = n^{-2} \sum_{i'}^i \sum_{j'}^i \sigma_{it'}^{\tau} \sigma_{jt'}^{\tau} f_{it'}^{\tau} f_{jt'}^{\tau} \leq n \sum_{j'}^i f_{jt'}^2 \sigma_{e_{jt'}^{\tau}}^2,$$

où n représente le nombre d'articles présents du groupe. La borne inférieure est obtenue dans le cas $f_{it}^{\tau} = 1/n$, et la borne supérieure dans le cas $\rho = 0$. Dans le premier cas, aucune correction de biais n'est nécessaire; dans le deuxième cas, nous prendrions $\hat{\Delta}(v) = \hat{v} - v$, où $\hat{v} =$

BIBLIOGRAPHIE

- ($t - t')$ $n \sum_{i'}^i f_{it'}^2 \hat{\sigma}_{e_{it'}^{\tau}}^2$ et $\hat{v} = \{\sum_{i'=1}^{i'} \gamma_2^* + \gamma_{i'}^* P_{i'-1}^{\tau-1}\} \hat{\sigma}_{e_{it'}^{\tau}}^2$. Ces valeurs correspondraient respectivement à $I_{t'}$ et à $I_{t'}$.
- ALDRICH, J. (1992). Probability and depreciation: a history of the stochastic approach to index numbers. *History of Political Economy*, 24, 657-87.
- BALK, B.M. (1980). A method for constructing price indexes for seasonal commodities. *Journal of the Royal Statistical Society, A*, 143, 68-75.
- BALK, B.M. (1995). Axiomatic price theory: a survey. *Revue Internationale de Statistique*, 63, 69-93.
- BRYAN, M.F., et CECCHETTI, S.G. (1993). The consumer price index as a measure of inflation. *Economic Review, Federal Reserve Bank of Cleveland*, 29, 15-24.
- CLEMENTS, K.W., et IZAN, H.Y. (1981). A note on estimating Divisia index numbers. *International Economic Review*, 22, 745-747.
- CLEMENTS, K.W., et IZAN, H.Y. (1987). The measurement of inflation: a stochastic approach. *Journal of Business and Economic Statistics*, 5, 339-350.
- DIEWERT, W.E. (1987). Index numbers. Dans *The New Palgrave: A Dictionary of Economics*, (Eds. J. Eatwell, M. Milgate, et P. Newman). London: MacMillan.
- DIEWERT, W.E. (1995). On the Stochastic Approach to Index Numbers. Document de discussion no. DP 95-31, Department of Economics, University of British Columbia.
- EICHHORN, W., et VOELTLER, J. (1976). *Theory of the Price Index*. Berlin: Springer-Verlag.
- FISHER, I. (1922). *The Making of Index Numbers*. Boston: Houghton Mifflin.
- HARVEY, A.C. (1990). *Forecasting, Structural Time Series Models and the Kalman Filter*. Cambridge: Cambridge University Press.
- KEYNES, J.M. (1930). *A Treatise on Money*. New York: Harcourt, Brace and Company.
- KOTT, P.S. (1984). A superpopulation approach to the design of price index estimators with small sampling biases. *Journal of Business and Economic Statistics*, 2, 83-90.
- SELVANATHAN, E.A., et RAO, D.S.P. (1994). *Index Numbers: A Stochastic Approach*. Ann Arbor: The University of Michigan Press.

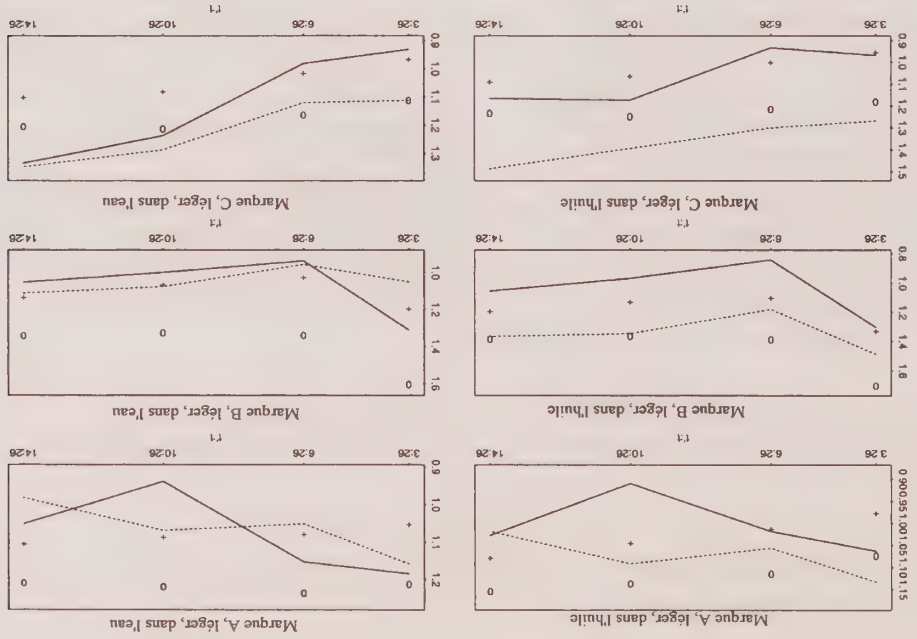


Figure 2. Quatre indices de prix pour quatre périodes de temps, thon léger

REMERCIEMENTS

L'auteur remercie B. Moulton, S. Scott, M. Reinsdorf, R. Tiller, B. Balk, et J. Aldrich pour des entretiens discussions portant sur les idées contenues dans la présente communication.

ANNEXE

Détails des équations (5) et (6).

Nous avons

$$G_{t'} = \prod_i \left(\frac{p_{t-1,i}}{p_{t,i}} \frac{p_{t-2,i}}{p_{t-1,i}} \dots \frac{p_{t',i}}{p_{t'+1,i}} \right)_{f'}$$
$$= \prod_i \left(r_{t-1,i}^{t'} \dots r_{t'+1,i}^{t'} \right)_{f'}$$

et lorsque

$$H_{t'} = \log(G_{t'}) = \sum_i f'_i \log(p_{t,i}^n / p_{t'+1,i}^n)$$

nous obtenons

$$H_{t'} = \sum_i f'_i \log(r_{t-1,i}^{t'} \dots r_{t'+1,i}^{t'})$$

$$= \sum_i^t f'_i (y_{t,i}^n + y_{t-1,i}^n + \dots + y_{t'+1,i}^n)$$

et également

avec

$$\gamma_{t,i}^{t'} \text{var}(\hat{y}_{t,i} | S'_i) = \left\{ \sum_i^{t'+1} \gamma_{t,i}^2 + \gamma_{t,i}^{t'} p_{t-1,i}^{t'} \right\} \sigma_{\varepsilon}^2$$

Nous considérons maintenant l'estimateur $\hat{I}_{t'} \equiv \exp(\hat{y}_{t,i} + \hat{y}_{t-1,i} + \dots + \hat{y}_{t'+1,i})$. Nous constatons que $E(\hat{I}_{t'}) = \exp(\mu_t + \mu_{t-1} + \dots + \mu_{t'+1} + 1/2 \hat{v})$, où $\hat{v} \equiv \text{var} \left(\sum_i^{t'+1} \hat{y}_{t,i} | S'_i \right) = \left\{ \sum_i^{t'+1} \gamma_{t,i}^2 \text{var}(\hat{y}_{t,i} | S'_i) + \right\}$ où $\sigma_{t'}^2$ représente la covariance de $\varepsilon_{t'}^n$ et $\varepsilon_{t'+1}^n$. Nous constatons que $v = (t - t') \sum_i^{t'+1} \sigma_{t'}^2 f'_i$, dans le cas particulier où les erreurs $\varepsilon_{t'}^n$ sont indépendantes et réparties de manière identique dans chaque période.

$$v = \text{var}(H_{t'}) \equiv \text{var}(H_{t'} | S'_i) = \text{var} \left(\sum_i^{t'+1} \sum_i^{t'+1} f'_i \varepsilon_{t,i}^n | S'_i \right) = (t - t') \sum_i^{t'+1} \sum_i^{t'+1} \sigma_{t'}^2 f'_i f'_i$$

et

$$E(H_{t'}) \equiv E(H_{t'} | S'_i) = \sum_i^t f'_i (\mu_t + \mu_{t-1} + \dots + \mu_{t'+1}) = \mu_t + \mu_{t-1} + \dots + \mu_{t'+1}$$

où les moments sont calculés en tenant compte de l'état $S_{t'}$ comme dans la section 4.3. Soit $v = \text{var}(H_{t'})$. Alors

$$I_{t'} = E(G_{t'}) = \exp(E(H_{t'})) + 1/2 \text{var}(H_{t'}))$$

Il varie, ce qui laisse penser que les indices classiques réagissent à un «bruit» présent dans les données et qu'en fait, très peu de variation a lieu au fond durant cette période de deux ans. On remarquera également, dans la figure 2, que le thon léger dans l'huile et le thon léger dans l'eau d'une même marque ont des comportements semblables, ce qui laisse penser que nous aurions peut-être dû prendre un groupe «homogène» plus large.

6. TRAVAUX ULTÉRIEURS

La recherche décrite dans la présente communication suggère plusieurs sujets pour des travaux de recherche ultérieurs.

Il faut notamment mettre au point des mesures de la précision et des estimations des indices fondés sur le modèle MMAB, sous forme de variances ou d'intervalles de confiance fondés sur le modèle à espace d'états. Même ceux qui doutent de la viabilité de l'application d'une méthode stochastique aux indices de prix trouvent la possibilité d'avoir une mesure de la précision attrayante (Dievert 1995). Il serait également utile d'obtenir des mesures de la précision d'autres indices standard, d'après le modèle à espace d'états.

Il est également souhaitable que du travail empirique soit effectué afin de déterminer avec plus de précision quels

La méthode fondée sur des modèles à espace d'états permet de composer avec l'absence de données (Harvey 1990, section 3.4.7). Une question importante que l'on se pose est celle de savoir dans quelle mesure ces modèles permettront de tenir compte de données manquantes dans l'estimation d'indices de prix. Étant donné que, dans la pratique, la plupart des données qui sont utilisées pour calculer des indices de prix portent sur un petit échantillon d'articles offerts, il faut connaître la robustesse des indices fondés sur des modèles à espace d'états à l'égard de l'absence de données.

Les algorithmes de lissage et de prévision des modèles à espace d'états sont bien connus. Leur utilisation dans la révision et la prévision d'indices pourrait être d'un grand intérêt. Enfin, dans la présente communication, nous avons abordé uniquement l'obtention d'un indice pour un seul groupe homogène. Il serait utile de créer un modèle à espace d'états qui réunirait plusieurs groupes et qui permettrait d'obtenir une mesure globale du pouvoir d'achat.

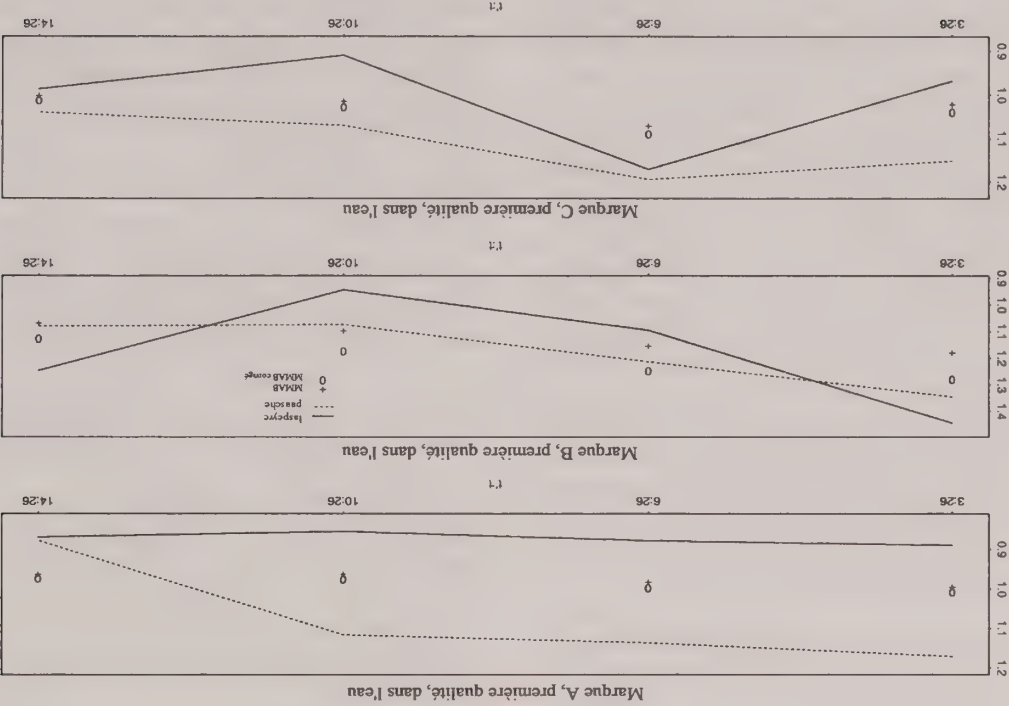


Figure 1. Quatre indices de prix pour quatre périodes de temps, thon de première qualité

groupe dans une micro-période tend à osciller autour du montant de l'augmentation (ou de la baisse) observée généralement au cours de la micro-période précédente. C'est une question d'observation courante: si la hausse des prix au cours d'un mois donné tend à être forte (ou faible), au cours du mois suivant, elle sera forte (ou faible), d'une manière correspondante. Etant donné que nous considérons un ensemble d'articles homogène, il est normal que leurs logarithmes de rapport de prix aient une moyenne commune. Nous renvoyons à des travaux ultérieurs la question de savoir comment intégrer des sous-indices à un indice global.

Le modèle (3) comporte le modèle (plus simple) unidimensionnel à marche aléatoire et bruit:

$$\bar{y}_t' = \mu_t' + \varepsilon_t', \quad \varepsilon_t' \sim N(0, \sigma_{\varepsilon}^2) \\ \mu_t' = \mu_{t-1}' + \eta_t', \quad \eta_t' \sim N(0, \sigma_{\eta}^2) \quad (4)$$

où $\bar{y}_t' = n^{-1} \mathbf{1}' y_t'$, $\varepsilon_t' = n^{-1} \mathbf{1}' \varepsilon_t'$, et $\sigma_{\varepsilon}^2 = n^{-1} \mathbf{1}' \sum_{t=1}^T \varepsilon_t' \varepsilon_t' \mathbf{1}$. Certains renseignements sont perdus lorsqu'on utilise l'équation (4); en revanche, la présupposition concernant la normalité est encore plus susceptible d'être juste. Pour plus de commodité, les calculs de l'étude décrite dans la section 5 étaient fondés sur le modèle unidimensionnel.

Le filtre de Kalman (Harvey 1990, section 3.2) peut être utilisé pour obtenir les estimations $\hat{\mu}_t'$, et $\hat{\sigma}_{\varepsilon}^2$, $\hat{\sigma}_{\eta}^2$ des paramètres d'état μ_t' et des variances σ_{ε}^2 , σ_{η}^2 respectivement. Nous définissons ensuite $I_{t'}' = E(G_{t'}' | S_t')$, où $G_{t'}' = \prod_t (D_{t'}'' / D_{t'}'')$ est une moyenne géométrique qui dépend des parts fixes $f_{t'}$, et où S_t' représente la totalité des paramètres d'état $\mu_{t'}$ au cours de la période t' , et également les «hyperparamètres» σ_{ε}^2 , σ_{η}^2 . Autrement dit, nous déterminons ce que nous considérons être le processus sous-jacent au cours de la période t' . Il s'ensuit que

$$I_{t'}' = \exp \left(\mu_{t'}' + \mu_{t-1}' + \dots + \mu_{t'+1}' + \frac{1}{2} v \right), \quad (5)$$

où $v = (t - t') \sum_{t'=t'+1}^t \sigma_{\eta}^2 f_{t'}' f_{t'}'$, avec σ_{η}^2 étant la covariance de ε_{η}'' et ε_{η}'' , sera habituellement d'un ordre inférieur à celui des paramètres d'état $\mu_{t'}''$. L'estimateur naturel de $I_{t'}'$ est $\hat{I}_{t'}' \equiv \exp(\hat{\mu}_{t'}' + \hat{\mu}_{t-1}' + \dots + \hat{\mu}_{t'+1}')'$; alors

$$E(\hat{I}_{t'}' | S_t') = \exp \left(\mu_{t'}' + \mu_{t-1}' + \dots + \mu_{t'+1}' + \frac{1}{2} v \right), \quad (6)$$

où v , donné dans l'annexe, n'est généralement pas égal à v , mais est souvent proche de cette valeur et, en tout cas, présente le même ordre de grandeur. La différence $\Delta(v) = \hat{v} - v$ peut être estimée par, disons, $\hat{\Delta}(v)$, ce qui donne un estimateur corrigé pour le biais $\hat{I}_{t'}' \equiv \hat{I}_{t'}' / \hat{\Delta}(v)$. Les expressions pour v et \hat{v} , ainsi qu'une proposition pour un $\hat{\Delta}(v)$ sont données dans l'annexe. On peut signaler que $\hat{\Delta}(v)$, et donc $\hat{I}_{t'}'$, dépend des poids $f_{t'}$, mais pas $I_{t'}'$.

5. ETUDE EMPIRIQUE

Afin de déterminer la faisabilité du calcul d'indices de prix à l'aide du modèle MMAB, et pour avoir une certaine idée du comportement de l'indice fondé sur ce modèle, on a effectué une petite étude empirique en utilisant des données sur les prix et les quantités du thon en boîte de l'Academic Database d'A.C. Nielsen. Le thon en boîte présente un comportement quelque peu instable en ce qui a trait aux prix et aux quantités, en raison de ventes fréquentes, qui s'effectuent parfois à des prix réduits de façon marquée.

Cette étude couvrait le nord-est des Etats-Unis et les 104 semaines des années 1992 et 1993. L'ensemble de données initial était plutôt volumineux. Afin de faciliter l'enquête, les données hebdomadaires ont été regroupées par périodes de quatre semaines, ce qui a donné 26 périodes réparties sur deux ans. Ainsi, aux fins de l'étude en question, les données qui ont été utilisées étaient des quantités cumulatives et des prix moyens pondérés d'après la quantité, pour des périodes de quatre semaines.

Les groupes homogènes ont été définis d'après la marque et le type de produit, de la façon suivante: 3 marques, désignées ici par les lettres A, B et C, de thon de «première qualité» dans l'eau, les trois mêmes marques de thon «léger» dans l'huile, et de nouveau les trois mêmes marques de thon «léger» dans l'eau, pour un total de 9 groupes.

L'étude était centrée sur 83 points de vente qui présentaient des quantités positives durant la plupart des périodes de 4 semaines, dans le cas de chacun des 9 groupes.

L'indice $I_{t'}'$ fondé sur le modèle MMAB et l'indice corrigé $\hat{I}_{t'}'$ fondé sur le même modèle ont été calculés pour quatre intervalles. Dans chaque cas, la période finale $t = 26$, et la période initiale a été prise successivement comme étant $t' = 3, 6, 10, 14$. A des fins de comparaison, nous avons également calculé les indices correspondants de Laspeyres et de Paasche. Ces deux indices standard servent également de base pour une comparaison indirecte avec les indices de Fisher et de Törnqvist, qui se situent environ à moitié chemin entre les deux indices précédents.

Les figures 1 et 2 (thon de première qualité et thon léger respectivement) donnent les valeurs des quatre indices pour les quatre intervalles, les points indiquant les indices d'espace d'états et les lignes les indices de Laspeyres et de Paasche. L'indice corrigé $\hat{I}_{t'}'$ est invariablenent plus élevé que l'indice non corrigé $I_{t'}'$. Il est à noter que puisque c'est la première période que nous faisons varier lorsque le tracé des indices est monotone vers le haut, cela semble indiquer une tendance à la baisse du prix du thon dans le groupe visé (et vice versa).

Nous constatons que les nouveaux indices ne sont pas aberrants par rapport aux indices classiques, se situant souvent entre les indices de Laspeyres et de Paasche, mais ils ont tendance à être nettement plus stables à mesure que

qu'une somme pondérée des tendances individuelles. Dievert propose la traduction suivante, sous forme d'un modèle, de l'objection de Keynes: étant donné que nous devons déterminer la tendance globale des prix par l'expression

$$\pi_t^* = \sum_{N=1}^t f_N \beta_N^*$$

le modèle (1) doit être remplacé par

$$(2) \quad \log \left(\frac{P_t}{P_t^{1,t}} \right) = \pi_t + \beta_t'' + \varepsilon_t''$$

où $\beta_t'' = \pi_t^* - \beta_t^{**}$ et $\sum_{N=1}^t f_N \beta_t'' = 0$. La différence déterminante entre cette équation et l'équation (1) est que, maintenant, les paramètres d'article β_t'' sont indexés en fonction du temps. Mais «alors, le modèle qui est obtenu comporte trop de paramètres à identifier». Ce fait devrait suffire à invalider la méthode.

Dievert (1995) n'aborde pas le modèle chronologique, beaucoup plus compliqué, de Bryan et Cechetti (1993). Parmi les communications précédentes, celle-ci est probablement celle qui se rapproche le plus de la nôtre. Elle porte sur l'utilisation d'un modèle complexe à espace d'états et du filtre de Kalman. À l'instar des autres communications examinées par Dievert, elle fait l'objet de l'objection de Keynes.

4. RÉEXAMEN DES INDICES DE PRIX

4.1 Présuppositions courantes

La modélisation stochastique du comportement des prix décrite dans la section précédente, qu'elle soit effectuée à l'aide des équations (1) ou (2), ou d'une équation similaire, présente trois caractéristiques dignes de mention; la modélisation est

1. complète: elle vise directement un «taux d'inflation» global qui comprend tous les articles;
2. atomistique: chaque article est modélisé individuellement et comporte son «propre» paramètre, son propre taux d'inflation $[\exp(\pi_t + \beta_t'), \text{ de manière distincte de tous les autres articles;}$
3. à périodes multiples: la modélisation pour la période comprise entre $t-1$ et t est distincte de celle s'appliquant à la période qui va de $t-2$ à $t-1$ etc.

C'est la combinaison de ces présuppositions qui donne lieu à l'argument de Dievert concernant le trop grand nombre de paramètres. La critique de Keynes vise avant tout la première caractéristique: un taux d'inflation global, pour lequel une augmentation ou une baisse du coût de la vie doit être un mélange pondéré de plusieurs tendances des

prix. On peut concéder cela sans aller jusqu'à inclure le point 2. Ce point est accepté tacitement dans le cas de presque toutes les constructions (non stochastiques) d'indices de prix. Cependant, il n'est pas évident du tout que chaque article a sa propre tendance en ce qui a trait au prix. En effet, il est probable que le prix d'articles différents (par exemple, la crème glacée de la marque X dans plusieurs supermarchés) a tendance à augmenter et à diminuer simultanément (du moins à long terme). Il existe des degrés d'homogénéité entre les articles. En tout cas, aucune de ces présuppositions n'est une composante nécessaire d'une approche stochastique à l'égard des indices de prix.

4.2 Un modèle élémentaire à espace d'états

Nous divisons la période comprise entre t' et t en sous-périodes $t', t' + 1, \dots, t-1, t$, et l'ensemble d'articles hétérogènes en sous-groupes homogènes, où la caractéristique qui définit l'homogénéité est la tendance à une similitude de comportement en ce qui a trait aux variations de prix. Nous présupposons deux choses:

1. $I_{t'}^{gr'}$ est un mélange d'indices $I_{t'}^{gr'}$ «homogènes»;
2. $I_{t'}^{gr'}$ peut être obtenu par enchaînement: $I_{t'}^{gr'} = \prod_{\tau=t'+1}^t I_{\tau}^{gr'-1,\tau}$ où $\tau = t' + 1, \dots, t$.

Nous nous concentrons sur un seul indice de groupe $I_{t'}^{gr'}$, et laissons tomber l'indice inférieur g , afin de simplifier la notation. Par conséquent, pour le reste de la présente communication, nous nous concentrons sur le «sous-indice» $I_{t'}^{gr'}$.

Nous allons maintenant créer un modèle élémentaire à espace d'états (Harvey 1990, chapitre 3) pour les logarithmes des rapports de prix à l'intérieur du groupe. Supposons que le groupe comprend un nombre n d'articles. Dans le cas de $i = 1, \dots, n$, soient $r''_i \equiv P''_i / P_{t-1,t}$ les rapports de prix relatifs à la micro-période, et $y''_i \equiv \log(P''_i) = \log(P''_i) - \log(P_{t-1,t})$, leurs logarithmes. La raison pour laquelle nous utilisons des logarithmes est que d'importants travaux empiriques, en premier lieu ceux d'Edgeworth (voir Dievert (1995)) laissent penser que les logarithmes des rapports de prix seront beaucoup plus susceptibles d'avoir une distribution normale que les rapports de prix eux-mêmes, qui peuvent être considérablement dissymétriques. La distribution normale des erreurs est une présupposition standard dans le cas des modèles à espace d'états. Soit $y'_i \equiv (y_{1i}, \dots, y_{mi})$ et $\mathbf{1}$ un vecteur de un de longueur m . Considérons le modèle multidimensionnel à marche aléatoire et bruit (MMAB):

$$(3) \quad y'_i = \mathbf{1} \mu'_i + \varepsilon'_i, \quad \varepsilon'_i \sim MVN(0, \Sigma^{\varepsilon})$$

$$\mu'_i = \mu_{i,t-1} + \eta_i, \quad \eta_i \sim N(0, \sigma^{\eta\eta})$$

où $\varepsilon'_i, \eta_i, \tau \in (t', t' + 1, \dots, t-1, t)$ sont mutuellement indépendants. Le modèle laisse entendre que le montant de l'augmentation (ou de la baisse) de l'ensemble des prix du

fixes, est le seul indice qui satisfait aux cinq axiomes relatifs aux indices de prix qui sont énoncés dans Balk (1995) ainsi qu'au «test de circularité», selon lequel $I_{t'} = I_{t''} I_{t''}^{t'}$ lorsque $t' < t < t''$. L'inversion du temps est une conséquence immédiate.

Les indices qui passent la plupart des tests sont généralement ceux qui intègrent des données quantitatives tirées des deux périodes; c'est le cas, par exemple, de l'indice de Fisher

$$F_{t',t''} = (I_{t',t''} P_{t',t''})^{1/2}$$

et de l'indice de Törnqvist

$$T_{t',t''} = \left(\prod_{i=1}^N \left(\frac{P_{t',t''}}{P_{t''}} \right)^{f_{t',t''}^i} \right)^{1/2}$$

où $f_{t',t''}^i = (f_{t',t''}^i + f_{t'',t'}^i)/2$. Souvent, on ne peut pratiquement pas distinguer l'un de l'autre les indices de Fisher et de Törnqvist. On peut trouver d'autres analyses de la méthode par tests dans Balk (1995), Diewert (1987), ainsi que dans Eichhorn et Voeller (1976).

La deuxième façon d'évaluer les formules relatives aux indices est la méthode «économique». Celle-ci définit un indice générique qui prend la forme suivante:

$$I_{t',t''} = \frac{C(p_{t',t''}, U)}{C(p_{t',t''}, U)}$$

où $U = U(q^1, \dots, q^N)$ est une «fonction utilitaire» bien définie et $C(p_{t',t''}, U)$ est le coût minimal aux prix $p_{t',t''}$, de l'atteinte du niveau de vie ou «utilité» U . Dans le cas d'une fonction utilitaire en particulier, on cherche à savoir si une formule donnée peut être considérée comme une bonne approximation de l'indice du coût de la vie correspondant. Comme la méthode des tests, cette deuxième façon de procéder à tendance à donner des indices qui comprennent des données quantitatives relatives aux deux périodes. Voir Diewert (1987) pour plus détails.

3. LA MÉTHODE STOCHASTIQUE

Aldrich (1992) présente l'historique des premiers essais de modélisation des rapports de prix, ou des logarithmes de rapports de prix, à l'aide d'un paramètre courant qui représente le taux global de croissance des prix. Une idée fondamentale contenue dans l'étude de cet auteur est que l'application de la méthode stochastique aux indices de prix, bien qu'elle soit un exemple d'une première application de la statistique à des questions économiques, est morte de sa belle mort. Diewert (1995) analyse lui aussi ces exemples, ainsi que des exemples plus récents de la modélisation statistique de rapports de prix. La difficulté que Diewert, à l'instar de Keynes (1930), voit dans ce genre de modélisation est illustrée par une modèle de Clements et Izan (1987).

La période comprise entre t' et t est divisée en segments de même durée, ce qui donne des intervalles relativement courts représentés de manière générique comme étant compris entre $t - 1$ et t . Le logarithme des rapports de prix pour une telle «micro-période» est donné par l'expression suivante:

$$(1) \quad \log \left(\frac{P_{t-1,t}}{P_t} \right) = \pi_t + \beta_t + \varepsilon_t$$

où $\varepsilon_t \sim (0, \sigma_t^2/f_t)$. Dans le modèle de ces auteurs, les f_t représentent la part de dépenses relative à l'article t , au cours de la période comprise entre t' et t . À des fins d'identification, on présume que $\sum_{i=1}^N f_i \beta_i = 0$. Ces présuppositions mènent à un estimateur du maximum de vraisemblance

$$\hat{\pi}_t = \sum_{i=1}^N f_i \log \left(\frac{P_t}{P_{t'}} \right)$$

ce qui donne un estimateur du maximum de vraisemblance (EMV) de la tendance des prix durant la période courte

$$\exp(\hat{\pi}_t) = \prod_{i=1}^N \left(\frac{P_t}{P_{t'}} \right)^{f_i};$$

On peut également obtenir des estimations de β_t et de σ_t^2 , ainsi que des estimations de la précision, par exemple, de la variance de $\hat{\pi}_t$. Ainsi, une nouvelle fondation statistique semble être placée sous un vieil estimateur. Diewert (1995) formule plusieurs objections, dont aucune ne peut être prise à la légère. L'objection principale est la suivante:

«...l'objection fondamentale de Keynes (Keynes 1930, p. 78): 'La variation du niveau des prix [$\exp(\pi_t)$] qui devrait avoir eu lieu s'il n'y avait pas eu de variation dans les prix relatifs, n'est plus pertinente si les prix relatifs ont effectivement subi une variation, car cette variation a eu elle-même une incidence sur le niveau des prix' » (Traduction).

Si, par exemple, le prix du pain varie par rapport au prix des voitures, cette même variation entraîne une variation du niveau global des prix.

L'objection de Keynes n'est pas tout à fait claire. Pourquoi ne peut-il y avoir deux aspects de la variation des prix, un aspect global et un aspect particulier? Cependant, il n'est pas difficile de reconnaître que les tendances des prix considérées individuellement sont primaires, une tendance des prix considérés globalement ne pouvant être

par Keynes et Diewert, ce qui mène naturellement à l'utilisation de modèles à espace d'états (section 4). Nous présentons notamment les résultats de l'application d'un modèle relativement simple à marche aléatoire et bruit à des données de balayage tirées de l'Academic Data Base d'A.C. Nielsen (section 5). Nous évaluons le nouvel indice dans la section 6, en mentionnant les travaux de recherche ultérieurs qui pourraient être utiles.

2. HISTORIQUE DE LA QUESTION

Par indice des prix à la consommation (IPC) on entend un nombre unique qui indique la variation du pouvoir d'achat des consommateurs entre une période t' et une période t . Les éléments de base bruts de cet indice sont les prix des divers articles que l'on peut se procurer (au moins aux deux époques

$$p^t = (p^{t1}, \dots, p^{tN}), \quad t = t', t$$

ainsi que les quantités des articles vendus

$$q^t = (q^{t1}, \dots, q^{tN}), \quad t = t', t.$$

(Cependant, dans la pratique, les données quantitatives relatives aux deux époques en question ne sont pas disponibles, et il faut alors recourir à des données de remplacement d'un type ou d'un autre). L'IPC est obtenu à l'aide d'une «formule» dans laquelle on utilise les éléments bruts suivants:

$$I_{t',t} = f(p^{t'}, p^t, q^{t'}, q^t),$$

où $f(\cdot)$ est une fonction de l'une des nombreuses expressions possibles. La plupart de ces expressions existent depuis longtemps et ont été décrites de façon approfondie dans les textes spécialisés qui portent sur les indices de prix.

À titre d'exemple, nous mentionnons ici l'indice de Laspeyres

$$L_{t',t} = \frac{\sum_{i=1}^N b^i p^{t'i}}{\sum_{i=1}^N b^i p^{t'i}} = \frac{\sum_{i=1}^N b^i p^{t'i}}{\sum_{i=1}^N b^i p^{t'i}},$$

où $f^{t'} = q^{t'} p^{t'}$, $f^t = q^t p^t$, les «rapports de prix», L'indice de Laspeyres fait appel aux quantités relatives à la période qui vient en premier comme base fixe pour la comparaison des prix antérieurs et des prix qui suivent. L'indice Laspeyres (ou une variante proche) a tendance à être l'indice le plus ciblé par les gouvernements, en raison de sa simplicité et de sa transparence aux yeux du profane.

qui étalonne les prix d'après les quantités de la période subséquente. La plupart des autres indices qui sont fondés sur d'autres formules se situent généralement entre les indices de Paasche et de Laspeyres.

À titre de référence ultérieure dans la présente communication, nous mentionnons un indice fondé sur la moyenne géométrique, où des poids non négatifs, fixes f_i , donnent au total 1:

$$P^{t',t} = \frac{\sum_{i=1}^N q^{t'i} p^{t'i}}{\sum_{i=1}^N q^{t'i} p^{t'i}}.$$

Le pendant naturel de l'indice de Laspeyres est l'indice de Paasche

On utilise le terme «moyenne géométrique» pour référer à cet indice.

Fisher (1922) analyse ces formules, ainsi que de nombreuses autres formules d'indices. Il introduit notamment l'approche dite «des tests» pour effectuer un choix parmi les diverses possibilités de formule $f(\cdot)$; dans le cas de cette méthode, on décrit les propriétés («tests») qu'un indice raisonnable devrait nécessiter, puis on détermine dans quelle mesure chaque formule d'indice satisfait à ces critères.

Un des tests est la condition de réversibilité dans le temps: $I_{t',t''} I_{t'',t} = 1$. Deux indices qui sont toujours utilisés dans le monde mais qui ne passent pas ce test sont l'indice Carli-Sauerbach $C_{t',t}^{t''} = \sum_{i=1}^N f_i p^{t'i} p^{t''i} / p^{t''i}$ et la moyenne géométrique $G_{t',t}^{t''} = \prod_{i=1}^N (p^{t'i} / p^{t''i})^{f_i}$ qui fait appel aux dépenses de la période antérieure plutôt qu'à des poids fixes. On peut montrer facilement que $C_{t',t}^{t''} > 1$, en utilisant l'inégalité de Cauchy-Schwarz, ce qui laisse penser que cet indice sera trop élevé.

Si une augmentation des prix de l'article i a tendance à produire une augmentation des dépenses, alors $G_{t',t}^{t''} < 1$, de sorte que dans de telles conditions, la moyenne géométrique fondée sur la période antérieure tend à être trop basse. D'autre part, si une augmentation des prix de l'article i tend à produire une diminution des dépenses, alors $G_{t',t}^{t''}$ donne des indications trop élevées. En général on peut donc s'attendre à ce que cet indice soit plutôt irrégulier.

Cette constatation suggère l'axiome suivant: les indices de prix qui sont exprimés sous la forme d'une moyenne géométrique ne devraient pas lier des poids aux prix observés dans une des périodes comparées; les indices qui prennent la forme d'une moyenne arithmétique ne devraient pas comporter des poids indépendants des prix en question. De façon contrastante par rapport à $G_{t',t}^{t''}$, la moyenne géométrique $G_{t',t}^{t''} = \prod_{i=1}^N (p^{t'i} / p^{t''i})^{f_i}$ qui comporte des poids

Les enquêtes sur les indices de prix en tant qu'études quasi-longitudinales

ALAN H. DORFMAN¹

RÉSUMÉ

Pour calculer les indices de prix, il faut recueillir des données relatives à un «même article» (en fait, un ensemble d'articles définis avec précision) durant diverses périodes. La question qu'on se pose est celle de savoir si de telles données «quasi-longitudinales» peuvent être modélisées de manière à expliquer ce qu'est un indice des prix. Des chercheurs de pointe spécialisés dans les questions relatives aux indices de prix ont émis des doutes quant à la possibilité d'utiliser la modélisation statistique pour caractériser de tels indices. Dans la présente communication, on propose un simple modèle à espace d'états relatif aux données sur les prix qui donne un indice des prix à la consommation exprimé d'après les paramètres du modèle.

MOTS CLÉS : Modèle à marche aléatoire et bruit; modèle à espace d'états; indice de Laspeyres; indice de Paasche; indice des prix géométrique.

1. INTRODUCTION

L'échantillonnage d'enquête utilisé pour calculer un indice des prix à la consommation comporte la surveillance d'un article donné au fil du temps, afin de déterminer ses prix à diverses époques. Seulement voilà, normalement, on ne suit pas exactement le même article (en effet, ce n'est pas le prix de la boîte de soupe aux tomates de marque Y que l'on trouve au point de vente Z qui est vérifié à plusieurs reprises, car il est probable que la boîte en question aura été vendue et consommée entre deux visites de la personne chargée de l'enquête), mais plutôt d'une succession d'articles qui correspondent tous à la même description («Boîte de 8 oz de soupe aux tomates avec harengs de marque Y, vendue au point de vente Z») et dont le prix est noté à diverses époques. Autrement dit, il s'agit essentiellement d'un groupe d'articles qui correspondent à une description précise suivie au fil du temps. C'est pourquoi les enquêtes qui portent sur les indices de prix peuvent être qualifiées de «quasi-longitudinales», par opposition aux enquêtes longitudinales, dans lesquelles on suit des articles distincts au fil du temps. Il est néanmoins raisonnable d'espérer que le fait d'effectuer des relevés à plusieurs reprises au fil du temps puisse mener à des méthodes d'estimation qui pourraient tirer davantage de l'aspect chronologique des enquêtes de ce type.

Compte tenu de cet espoir, la présente communication se penche sur une question qui a été généralement ignorée par les statisticiens et les économistes, ou qui tout au plus a reçu une réponse négative. Cette question est la suivante: est-ce qu'on peut traiter un indice des prix à la consommation (IPC) d'un point de vue statistique? Autrement dit, est-ce que le paramètre qui caractérise la «variation du coût de la vie» d'une période à une autre, et que les enquêtes sur les

indices de prix essaient d'estimer, peut être défini d'après un modèle stochastique?

Aldrich (1992) donne une interprétation historique des premières tentatives, entreprises par Jevons, et surtout par Edgeworth, d'intégrer des hypothèses distributionnelles aux indices des prix à la consommation. Les communications récentes ayant pour objet la modélisation stochastique sont celles de Balk (1980), Clements et Izan (1981, 1987), Bryan et Cecchetti (1993), Kott (1984), et Selvanathan et Rao (1994). Dievert (1995) passe en revue et critique ces tentatives en invoquant un argument de Keynes (1930) comme motif déterminant pour rejeter l'approche stochastique.

Dans la présente communication, nous proposons une approche précise en ce qui a trait à la modélisation de l'indice des prix à l'aide de modèles à espace d'états, et nous soumettons provisoirement un tel modèle. Celui-ci est appliqué à des données de balayage afin de montrer la faisabilité d'un indice fondé sur ces mêmes données. La méthode que nous considérons contourne la critique de Keynes d'une façon fondamentale et offre la perspective des nombreux avantages que peut apporter une solide modélisation statistique, y compris, peut-être, une simplification du processus d'échantillonnage d'enquête.

Ci-après, nous allons d'abord passer brièvement en revue la définition d'un indice des prix ainsi que les deux méthodes (non stochastiques) qui sous-tendent de manière prédominante le choix des indices (section 2). Nous examinons notamment l'exemple de modèle statistique pour indices des prix qui est présenté par Bryan et Cecchetti (1993), ainsi que la formulation de l'objection de Keynes par Dievert (section 3). Nous présentons par la suite une méthode de modélisation d'un indice des prix à la consommation qui contourne les difficultés mentionnées

¹ Alan H. Dorfman, U.S. Bureau of Labor Statistics, Room 4915, 2 Massachusetts Ave. N.E., Washington, D.C., 20212-0001, U.S.A.; courriel électronique: dorfman_a@bls.gov.

estimations plus efficaces des paramètres marginaux et de la variation, comparativement à la régression logistique ordinaire et à l'estimation simple de la variation. Enfin, nous avons décrit la possibilité qu'offre la régression logistique multidimensionnelle de représenter une structure de dépendance complexe à l'aide d'un petit nombre de paramètres.

En utilisant les résultats obtenus par Glonek et McCullagh (1995), il est possible d'appliquer les exemples que nous avons présentés ici à des réponses multidimensionnelles de type nominal ou ordinal, avec des variables explicatives discrètes ou continues. La méthode que nous avons décrite permet également de tenir compte des poids d'échantillonnage (Salamon 1998). Dans le cas de l'ESPA, nous avons constaté que les poids d'échantillonnage ont peu d'incidence sur les estimations de paramètres du modèle logistique multidimensionnel. L'erreur-type des estimations de paramètres était gonflée d'environ 15 %. Cette augmentation modérée de la variabilité de ces estimations, qui est due aux poids d'échantillonnage, est plausible. En effet, étant donné qu'on choisit une seule personne par ménage dans le cadre de l'ESPA, nous ne nous attendions pas à un effet de grappe important.

Outre que pour les types d'analyse que nous avons présentés ici, la régression logistique multidimensionnelle peut être utilisée également pour modéliser des probabilités de non-réponse dans le cas d'études longitudinales. De tels modèles pourraient être utiles lorsqu'il faut corriger des poids d'échantillonnage pour tenir compte de la non-réponse. La possibilité qu'offre la régression logistique multidimensionnelle d'obtenir un modèle parcimonieux des données pourrait avoir de l'intérêt dans le cas de l'estimation relative à de petites régions. Des estimateurs pour une région géographique donnée pourraient notamment être fondés sur des modèles relatifs à une région plus vaste choisie de manière appropriée.

Bien que nous n'ayons pas rencontré des difficultés importantes dans les exemples que nous avons présentés ici, des études ultérieures pourraient être nécessaires en ce qui a trait au problème des tableaux claisés. L'inversion de la transformation logistique multidimensionnelle est une opération importante lorsqu'il y a un grand nombre de cellules vides. La méthode proposée par Lang (1996), dans laquelle l'inversion de la fonction de liaison est évitée en spécifiant les modèles par l'entremise de contraintes, pourrait être intéressante dans un tel contexte. Un autre aspect à étudier est l'incidence des erreurs de classification sur les estimations de paramètres du modèle logistique multidimensionnel.

REMERCIEMENTS

Pour la rédaction du présent article, nous avons bénéficié d'entretiens avec des collègues de l'Office fédéral de la

statistique, parmi lesquels méritent d'être mentionnés spécialement Beat Hultiger et Philippe Bichenberger. L'aide d'Ariane Bender, de la section responsable de l'Enquête suisse sur la population active, a été extrêmement utile. L'auteur remercie également le rédacteur et deux examinateurs pour leurs excellents conseils, qui ont permis d'améliorer cet article de manière importante.

BIBLIOGRAPHIE

BINDER, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *Revue Internationale de Statistique*, 51, 279-292.

DIGGLE, P.J., LIANG, K.-Y., et ZEGGER, S.L. (1994). *Analysis of Longitudinal Data*. Oxford: Oxford University Press.

GERFIN, M. (1996). *Entwicklung von ökonomischen Modellen zur Analyse der Dynamik auf dem schweizerischen Arbeitsmarkt*. SLPs-News, Swiss Federal Statistical Office, Berne.

GLONEK, G.F.V., et McCULLAGH, P. (1994). Multivariate Logistic Models. Rapport technique 94-31, School of Information Science and Technology, Flinders University of South Australia, Adelaide.

GLONEK, G.F.V., et McCULLAGH, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society*, B, 57, 533-546.

HULTIGER, B., RIES, A., COMMENT, T., et BENDER, A. (1997). Weighing the Swiss Labour Force Survey. (Eds. C. Malagueira, S. Morgensthaler et E. Ronchetti). Dans *Conference on Statistical Science Honoring the Bicentennial of Stefano Franscini's Birth, Monte Verità, Switzerland*, Basel: Birkhäuser Verlag.

LANCASTER, T. (1990). *The Econometric Analysis of Transition Data*. Cambridge: Cambridge University Press.

LANG, J.B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Annals of Statistics*, 24, 726-752.

LIANG, K.-Y., ZEGGER, S.L., et QAOISH, B. (1992). Multivariate Regression Analysis for Categorical Data. *Journal of the Royal Statistical Society*, B, 54, 3-40.

McCULLAGH, P., et NELDER, J.A. (1989). *Generalized Linear Models*, (2ième édition). London: Chapman and Hall.

PFEFFERMAN, D. (1993). The role of sampling weights when modeling survey data. *Revue Internationale de Statistique*, 61, 317-337.

PFEFFERMAN, D., SKINNER, C., et KEITH, H. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society*, A, 161, Part 1, 13-32.

SALAMON, P.-A. (1998). Multivariate logistic regression for data from complex surveys. À paraître *Recueil: Symposium '98, Analyse longitudinale pour les enquêtes complexes*, Statistique Canada, Mai 1998.

ZEGGER, S.L., et LIANG, K.-Y. (1992). An overview of methods for the analysis of longitudinal data. *Statistics in Medicine*, 11, 1825-1839.

Tableau 4
Nombre de paramètres et valeur de la fonction de vraisemblance logarithmique dans le cas des estimations du maximum de vraisemblance

Modèle	Nombre de paramètres d'ordre				Vraisemblance
	1	2	3	4	
Modèle complet	20	20	10	2	-5 342,7
1	4	20	10	2	-5 345,4
2	4	20	0	0	-5 349,4
3	4	6	0	0	-5 365,2
4	4	3	0	0	-5 368,9
5	4	2	0	0	-5 369,5
6	4	0	0	0	-7 815,3

5. COMPARAISON AVEC L'ESTIMATION SIMPLE DE LA VARIATION

Dans la présente section, nous nous concentrons sur l'estimation de la différence, observée d'une année donnée à une autre, dans les probabilités d'occuper un emploi. Nous montrons que des estimations fondées sur la régression logistique multidimensionnelle sont plus efficaces que des estimations simples définies comme étant la différence existant entre les proportions de personnes qui occupent un emploi. Le modèle qui est considéré ici est le modèle 2 de la section 4, le sexe étant ajouté comme variable explicative supplémentaire. Dans le cas de chacun des sexes, nous avons un paramètre pour chacun des logits marginaux qui correspond à une année donnée. La dépendance longitudinale est prise en considération au moyen d'un modèle saturé pour les risques relatifs logarithmiques. Les paramètres de troisième et quatrième ordre sont établis à zéro. Ce modèle comporte donc 8 paramètres pour les logits marginaux et 40 paramètres pour les risques relatifs logarithmiques (2 sexes × 20 risques relatifs à l'intérieur des périodes d'observation; voir tableau 3). En inversant la transformation logistique multidimensionnelle, on peut également calculer les estimations des probabilités d'occuper un emploi et de la variation de ces probabilités d'une année à l'autre.

Un estimateur simple de variation est donné par la différence existant entre les proportions de personnes occupant un emploi qui est observée d'une année donnée à une autre. La variance, qui tient compte du chevauchement des deux échantillons, est donnée par l'expression suivante:

$$\frac{1}{n+r} \pi_{1+} (1 - \pi_{1+}) + \frac{1}{n+c} \pi_{+1} (1 - \pi_{+1}) - 2 \frac{n}{n} \frac{(n+r)(n+c)}{(\pi_{11} - \pi_{1+} \pi_{+1})},$$

où n est le nombre de cas au sujet desquels on dispose d'observations pour les deux années, r et c le nombre de cas au sujet desquels on dispose d'observations relatives à un an, π_{1+} la probabilité d'occuper un emploi durant les deux ans, et π_{+1} les probabilités marginales d'occuper un emploi.

Concernant les données de l'ESPA de la section 2, le tableau 5 indique les estimations de la différence de la probabilité d'occuper un emploi obtenues avec les deux méthodes. On remarquera que ces méthodes donnent des estimations de la variation similaires. Les erreurs-types des estimations simples sont, en moyenne, 30 % plus importantes que dans le cas de la régression logistique multidimensionnelle. L'efficacité relative moyenne de la régression par rapport aux estimations logistiques multidimensionnelles est de 1,7. Par comparaison, l'efficacité relative moyenne de la régression logistique multidimensionnelle par rapport à la régression logistique ordinaire est de 3,2.

Tableau 5
Variation de la probabilité d'occuper un emploi, canton de Vaud, 1992-1995

	Régression logistique multidimensionnelle		Comparaison	Estimation simple	
	Femme	Homme		Femme	Homme
92 vs. 93	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0031 (0,0107)	0,0283 (0,0116)
92 vs. 94	0,0168 (0,0134)	0,0184 (0,0102)	0,0356 (0,0149)	0,0219 (0,0128)	0,0245 (0,0102)
92 vs. 95	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0219 (0,0128)	0,0245 (0,0102)
93 vs. 94	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0219 (0,0128)	0,0245 (0,0102)
93 vs. 95	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0219 (0,0128)	0,0245 (0,0102)
94 vs. 95	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0219 (0,0128)	0,0245 (0,0102)
94 vs. 96	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0219 (0,0128)	0,0245 (0,0102)
95 vs. 96	0,0138 (0,0090)	0,0184 (0,0102)	0,0375 (0,0109)	0,0219 (0,0128)	0,0245 (0,0102)

6. CONCLUSIONS

Les analyses des données de l'ESPA que nous avons décrites dans le présent article montrent l'utilité de la régression logistique multidimensionnelle. La modélisation de la dépendance longitudinale est nécessaire pour obtenir un ajustement satisfaisant des probabilités relatives aux profils de réponse qui sont observés. En ne tenant pas compte de la dépendance longitudinale, nous obtenons des estimations ponctuelles acceptables des logits marginaux, mais les informations sur la structure détaillée des données sont perdues. La modélisation de la dépendance longitudinale permet également d'obtenir des

diminue le risque d'avoir des tableaux clairsemés lorsqu'on considère des périodes d'observation prolongées comportant un plus grand nombre de covariables.

Les modèles 3, 4 et 5 montrent qu'il serait néanmoins possible de simplifier grandement la description de la dépendance longitudinale sans perdre trop d'informations. En passant du modèle 2 au modèle 5, nous constatons que l'écart par rapport au modèle entièrement saturé n'augmente pas beaucoup (voir tableau 4). En outre, un examen des

résidus montre que les modèles 3, 4 et 5 correspondent aux données presque aussi bien que le modèle 2. D'autre part, si le modèle 2 requiert vingt paramètres pour décrire la dépendance longitudinale, le modèle 5 n'en nécessite que deux. Ce fait doit être opposé au modèle 6, qui présuppose l'indépendance entre les observations effectuées à des époques différentes; la vraisemblance logarithmique est beaucoup plus faible que dans le cas du modèle entièrement saturé (voir tableau 4) et l'ajustement aux données laisse à désirer.

Tableau 3

Estimations de paramètres et erreurs-types

Paramètres	Période	Modèle 2	Modèle 3	Modèle 4	Modèle 5	Modèle 6
logit 92	0,6348 (0,0350)	0,6360 (0,0352)	0,6348 (0,0352)	0,6347 (0,0352)	0,6471 (0,0409)	
logit 93	0,5555 (0,0335)	0,5570 (0,0338)	0,5597 (0,0335)	0,5601 (0,0335)	0,5509 (0,0396)	
logit 94	0,5440 (0,0324)	0,5407 (0,0325)	0,5402 (0,0326)	0,5397 (0,0325)	0,5377 (0,0374)	
logit 95	0,4699 (0,0317)	0,4711 (0,0320)	0,4710 (0,0320)	0,4712 (0,0320)	0,4705 (0,0351)	
β_{23}	(1) 23 4,2563 (0,3311)	4,2579 (0,1465)				
	(1) 234 4,2003 (0,2894)					
	(1) 2345 4,0859 (0,2954)					
	2345 4,4830 (0,2841)					
β_{34}	(1) 234 4,0894 (0,2794)	4,1111 (0,1310)				
	(1) 2345 3,9611 (0,2840)					
	2345 4,0989 (0,2600)					
	345 4,2490 (0,2468)					
β_{45}	(1) 2345 5,3992 (0,3854)	4,5561 (0,1389)				
	2345 3,9779 (0,2544)					
	345 4,7288 (0,2735)					
	45 4,5069 (0,2600)					
β_{24}	(1) 234 3,7168 (0,2641)	3,8371 (0,1442)				
	(1) 2345 4,2560 (0,3059)					
	2345 3,5330 (0,2370)					
	(1) 2345 4,4000 (0,3098)	3,7913 (0,1334)				
	2345 3,6493 (0,2396)					
	345 3,6116 (0,2192)					
β_{25}	(1) 2345 4,3984 (0,3173)	3,5774 (0,1530)				
	2345 3,2209 (0,2256)					
γ_1		4,3260 (0,0928)				
γ_2		3,8519 (0,1050)				
γ_3		3,5340 (0,1495)				
δ		4,7341 (0,1266)				
γ		-0,4191 (0,0653)				

question.

LONGITUDINALE

d'un petit nombre de paramètres.

Nous considérons six modèles à complexité décroissante (voir tableau 2). Dans le cas de tous les modèles, nous avons un paramètre pour chacun des logits marginaux qui correspond à une époque d'observation donnée. Ce paramètre est désigné par la notation $\eta_i = \beta_i$. Étant donné que les époques d'observation sont les deuxièmes trimestres des années 1992 à 1995, nous prenons $i = 2, 3, 4, 5$. Ainsi, β_j , par exemple, correspond au logit de la probabilité d'occuper un emploi en 1993. De façon analogue, les indices pour les paramètres d'ordre supérieur vont de 2 à 5. Dans le cas du modèle 1, nous prenons un modèle saturé pour la dépendance longitudinale, c'est-à-dire que nous avons un paramètre pour chacune des interactions d'ordre 2, 3 ou 4 à l'intérieur de chaque période d'observation. Dans le cas des modèles 2 à 5, nous supposons que les interactions d'ordre 3 et 4 sont toutes égales à zéro. La dépendance longitudinale est alors décrite seulement sous forme de risques relatifs logarithmiques. Dans le cas du modèle 2, nous avons recours à un modèle saturé pour les risques relatifs logarithmiques. Dans le modèle 3, nous laissons tomber la covariable période d'observation, c'est-à-dire que nous supposons que les risques relatifs logarithmiques sont les mêmes pour toutes les périodes d'observation. Dans le modèle 4, nous utilisons des risques relatifs logarithmiques qui dépendent uniquement de la différence entre les époques d'observation. Il est à noter que dans le modèle 4, le paramètre γ_1 correspond à la contrainte $\beta_{23} = \beta_{34} = \beta_{45} = \beta_{45}$ à laquelle les paramètres du modèle 3, et la même chose vaut pour γ_2 et γ_3 . Dans le cas du modèle 5, on suppose un modèle linéaire pour les risques relatifs logarithmiques stationnaires. Enfin, dans le modèle 6, nous supposons que les observations effectuées à des époques différentes sont indépendantes. Il est à noter que dans ce cas, la régression logistique multimensionnelle équivaut à la régression logistique ordinaire.

Tableau 2

Paramètres	
Logits	Paramètres de 3 ^{ème} et 4 ^{ème} ordre
Risques relatifs	
marginaux	
Modèle	
1	$\eta_i = \beta_{fj, \text{période}}$ $\eta_{ijk} = \beta_{fjk, \text{période}}$ $\eta_{ijkl} = \beta_{fjkl, \text{période}}$
2	$\eta_i = \beta_i$ $\eta_{ij} = \beta_{fj, \text{période}}$ $\eta_{ijk} = 0, \eta_{ijkl} = 0$
3	$\eta_i = \beta_i$ $\eta_{ij} = \beta_{fj}$ $\eta_{ijk} = 0, \eta_{ijkl} = 0$
4	$\eta_i = \beta_i$ $\eta_{ij} = \gamma_i - i - j $ $\eta_{ijk} = 0, \eta_{ijkl} = 0$
5	$\eta_i = \beta_i$ $\eta_{ij} = \delta + \lambda \cdot i - j $ $\eta_{ijk} = 0, \eta_{ijkl} = 0$
6	$\eta_i = \beta_i$ $\eta_{ij} = 0$ $\eta_{ijk} = 0, \eta_{ijkl} = 0$

logarithmique pour le modèle entièrement saturé.

Dans l'ensemble, nous constatons que la forme pré-supposée de la dépendance longitudinale semble avoir peu d'incidence sur les estimations des logits marginaux. Il s'agit là d'une caractéristique souhaitable, étant donné que les logits marginaux seraient normalement les paramètres auxquels on s'intéresse. Les erreurs-types des logits marginaux sont presque les mêmes dans le cas des modèles qui tiennent compte de la dépendance longitudinale, mais ils sont gonflés dans une mesure d'environ 15 % dans le cas de la régression logistique ordinaire (modèle 6). On peut également montrer que les estimations des logits marginaux présentent entre elles une corrélation positive dans le cas de modèles qui présupposent une dépendance longitudinale, et aucune corrélation dans le cas de la régression logistique ordinaire. Dans l'exemple qui nous intéresse, nous avons constaté que la corrélation se situe entre 0,4 et 0,8. Ainsi, la modélisation de la dépendance longitudinale permet également d'obtenir des estimations plus efficaces de la différence existant entre les logits marginaux.

D'après l'ajustement du modèle 1, on peut voir que les paramètres d'interaction d'ordre 3 et 4 ne sont pas très différents de 0. Cela laisse penser que la dépendance longitudinale peut être décrite uniquement par les risques relatifs logarithmiques. Cette hypothèse est corroborée par l'écart incremental du modèle 2 par rapport au modèle 1, qui est de 7,9 pour 12 degrés de liberté. En outre, tous les paramètres du modèle 2 sont différents de zéro de manière importante, et un examen des résidus normalisés dans le cas des probabilités ajustées relatives aux profils de réponse ne révèle aucune anomalie. Dans le cas d'applications relatives à des statistiques officielles, le modèle 2 serait le modèle privilégié, étant donné qu'il est fondé sur un nombre de présuppositions aussi faible que possible tout en permettant une réduction importante du nombre de paramètres, ce qui

$$\eta_{123} = \log RR(Y_1, Y_2 | Y_3 = 1) - \log RR(Y_1, Y_2 | Y_3 = 2) \\ = \log \frac{\pi_{111}\pi_{221}}{\pi_{112}\pi_{222}} - \log \frac{\pi_{121}\pi_{211}}{\pi_{122}\pi_{212}}.$$

Lorsque $d = 4$, $\pi = (\pi_{1111}, \pi_{1112}, \dots, \pi_{2221}, \pi_{2222})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_3, \eta_{13}, \eta_{23}, \eta_{123},$$

$$\eta_4, \eta_{14}, \eta_{24}, \eta_{124}, \eta_{34}, \eta_{134}, \eta_{234}, \eta_{1234})^T.$$

Les paramètres η_i, η_{ij} et η_{ijk} , où $1 \leq i < j < k \leq 4$, sont définis comme ci-dessus à l'aide des tableaux marginaux appropriés. Le paramètre η_{1234} est un contraste des risques relatifs logarithmiques qui est donné par

$$\eta_{1234} = \log RR(Y_1, Y_2 | Y_3 = 1, Y_4 = 1)$$

$$- \log RR(Y_1, Y_2 | Y_3 = 1, Y_4 = 2)$$

$$- \log RR(Y_1, Y_2 | Y_3 = 2, Y_4 = 1)$$

$$+ \log RR(Y_1, Y_2 | Y_3 = 2, Y_4 = 2).$$

Une opération clé de l'estimation du maximum de vraisemblance est le calcul de l'inverse de la transformation logarithmique multidiimensionnelle. Pour faire en sorte que $\pi > 0$, nous utilisons $\pi = \exp v$, c'est-à-dire que nous cherchons à trouver v dans l'équation $\eta = C^T \log(L \exp v)$. En général, on ne dispose pas d'une solution explicite, de sorte qu'il faut utiliser une méthode itérative. On peut notamment appliquer les itérations de Newton-Raphson, comme nous l'expliquons ci-après. Pour plus de clarté, nous définissons les deux fonctions $\phi(\pi) = C^T \log(L\pi)$ et $\psi(v) = \phi(\exp v)$.

i) Commencer avec une approximation initiale v_0 .
ii) Prendre ensuite $v_{k+1} = v_k - [D\psi(v_k)]^{-1}(\phi(\exp v_k) - \eta)$, où $D\psi(v)$ est la matrice jacobienne de la fonction $\psi(v)$, et itérer jusqu'à convergence.

Les matrices jacobienes des fonctions $\phi(\pi)$ et $\psi(v)$ sont données respectivement par $D\phi(\pi) = C^T(\text{diag } L\pi)^{-1}L$ et $D\psi(v) = (\exp v) \cdot \text{diag}(\exp v)$.

3.2 Estimation du maximum de vraisemblance

Dans le cas d'une variable de réponse binaire observée à d époques il y a $q = 2^d$ profils de réponse possibles, où chaque profil $i = (i_1, \dots, i_d)$ ont comme valeur soit 1, soit 2. Dans le cas de indicatrice Y_{i_1, \dots, i_d} , qui est égale à 1 si l'on a observé le profil i , et à 0 dans le cas contraire. On obtient alors

$$P(Y_{i_1, \dots, i_d} = 1) = P(Y_1 = i_1, \dots, Y_d = i_d) = \pi_{i_1, \dots, i_d}.$$

nous pouvons déduire que

$$\mathcal{S}(\beta) = \mathcal{S}(\beta, X) = D\pi(\beta)^T(\text{diag } \pi(\beta))^{-1}D\pi(\beta).$$

Si nous avons un nombre n d'observations indépendantes $Y_k \sim M(1, \pi_k)$, $k = 1, \dots, n$, où $\eta_k = C^T \log(L\pi_k) = X_k^T \beta$, le vecteur de valeur numérique et la matrice d'information sont donnés alors par $s(\beta) = \sum_{k=1}^n s(\beta, Y_k, X_k)$ et par $\mathcal{S}(\beta) = \sum_{k=1}^n \mathcal{S}(\beta, X_k)$.

L'estimateur du maximum de vraisemblance de β est la solution de $s(\beta) = 0$, qui peut être trouvée à l'aide de l'algorithme de cotation de Fisher qui, à partir d'une quelconque valeur initiale β_0 , itère la séquence $\beta_{m+1} = \beta_m + \mathcal{S}_m^{-1}(\beta_m)s(\beta_m)$ jusqu'à convergence.

Des profils de réponse incomplets peuvent être intégrés facilement à l'analyse. En particulier, lorsqu'un quelconque sous-ensemble des variables de réponse X_1, X_2, \dots, X_c est enregistré pour une unité donnée, la distribution des probabilités sur ce tableau marginal de grandeurs c est multinomiale et, en raison de la reproductibilité de la transformation logarithmique multidiimensionnelle, un modèle de régression logarithmique multidiimensionnelle s'applique au tableau des probabilités. En outre, la matrice de plan

probabilités sont données par le vecteur π .

Les modèles de régression logarithmique multidiimensionnelle sont donc définis pour être ceux de la forme $\eta = X\beta$, où X est une matrice $q \times p$ de variables explicatives, β un vecteur de grandeur p de paramètres inconnus, et où $\eta = C^T \log(L\pi) = \phi(\pi)$.

Si nous prenons y comme observation du vecteur aléatoire X , nous pouvons alors exprimer le noyau de la fonction de vraisemblance logarithmique comme suit: $l(\beta; y) = y^T \log \pi(\beta)$, où, en utilisant l'inverse de la transformation logarithmique multidiimensionnelle, nous pouvons exprimer les probabilités conjointes π comme une fonction du paramètre inconnu β , sous la forme $\pi(\beta) = \phi^{-1}(X\beta)$. Le vecteur de valeur numérique est donné par

$$s(\beta) = s(\beta, y, X) = D\pi(\beta)^T(\text{diag } \pi(\beta))^{-1}y,$$

où $D\pi(\beta)$, c'est-à-dire la matrice jacobienne de la fonction $\pi(\beta)$ qui lie le paramètre β au vecteur de probabilités π , est donnée par $D\pi(\beta) = [D\phi(\phi^{-1}(X\beta))]^{-1}X$, et où $D\phi(\pi) = C^T(\text{diag } L\pi)^{-1}L$ est la matrice jacobienne de la fonction de liaison. La matrice d'information est définie comme suit: $\mathcal{S}(\beta) = E[s(\beta)s(\beta)^T]$. Il s'ensuit, d'après la présupposition relative à la distribution de X , que $E(XY^T) = \text{diag } \pi$, d'où

où $1_2^T = (1, 1)$, L la matrice unité deux par deux et $\tilde{C} = (1, -1)^T$ (Glonek et McCullagh 1994). À titre d'exemple, nous considérons des périodes d'observation d'une durée $d = 1, 2, 3, 4$. Dans le cas de $d = 1$, $\pi = (\pi_1, \pi_2)^T$ et $\eta = (\eta_0, \eta_1)^T = (\log \pi_+, \log Y_1^T)$, où l'indice positif indique la sommation et où $\log Y_1$ est défini comme suit:

$$\log Y_1 = \log \frac{P(Y_1 = 1)}{P(Y_1 = 2)} = \log \frac{\pi_1}{\pi_1 - \pi_2} = \log \frac{\pi_2}{\pi_1}.$$

Dans ce cas, la transformation logistiqua multidimensionnelle est équivalente à la transformation logistiqua habituelle. Il est à noter que même si le paramètre $\eta_0 = \log \pi_+ = 0$ est tout à fait superflu, il est utile de le conserver comme moyen permettant de s'assurer que l'application $\pi \mapsto \eta$ est de plein niveau et exprime également l'exigence selon laquelle $\pi_+ = 1$.

Lorsque $d = 2$, $\pi = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12})^T =$$

$$(\log \pi_{++}, \log Y_1, \log Y_2, \log OR(Y_1, Y_2))^T$$

où

$$RR(Y_1, Y_2) =$$

$$\frac{P(Y_1 = 1, Y_2 = 1)P(Y_1 = 2, Y_2 = 2)}{P(Y_1 = 1, Y_2 = 2)P(Y_1 = 2, Y_2 = 1)} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

est le risque relatif, un grandeur qui mesure l'association entre les variables Y_1 et Y_2 . Les paramètres η_1 et η_2 sont les logits marginaux aux époques t_1 et t_2 ; par exemple:

$$\eta_1 = \log Y_1 = \log \frac{\pi_{1+}}{(1 - \pi_{1+})}.$$

Lorsque $d = 3$, $\pi = (\pi_{111}, \pi_{112}, \pi_{121}, \pi_{222})^T$ et

$$\eta = (\eta_0, \eta_1, \eta_2, \eta_{12}, \eta_{13}, \eta_{23}, \eta_{123})^T.$$

Les paramètres η_1, η_2 et η_3 sont les logits marginaux aux époques t_1, t_2 et t_3 . Les paramètres η_{12}, η_{13} et η_{23} sont les risques relatifs logarithmiques des tableaux marginaux bidimensionnels correspondants; par exemple:

$$\eta_{23} = \log RR(Y_2, Y_3) = \log \frac{\pi_{+11}\pi_{+22}}{\pi_{+12}\pi_{+21}}.$$

Le paramètre η_{123} est un contraste des risques relatifs qui est donné par

Glonek et McCullagh (1995). Nous avons effectué certaines comparaisons entre la régression logistiqua multidimensionnelle et le logiciel PROC GENMOD de SAS (version 6.12). Ce logiciel permet d'ajuster des modèles de réponses corrélées d'après la méthode des EBG. Nous avons obtenu des estimations des logits marginaux très semblables. La méthode des EBG s'est avérée légèrement moins efficace que la régression logistiqua multidimensionnelle. Un des inconvénients de cette méthode est qu'elle ne permet pas d'obtenir des estimations de toutes les probabilités marginales. Le modèle logistiqua multidimensionnel ne présente pas cet inconvénient parce que ses paramètres sont estimés selon la méthode du maximum de vraisemblance.

Soient Y_1, Y_2, \dots, Y_d un nombre d d'observations répétées de la même variable binaire, effectuées aux époques $t_1 < t_2 < \dots < t_d$, et soit

$$\pi_{t_1, t_2, \dots, t_d} = P(Y_1 = t_1, Y_2 = t_2, \dots, Y_d = t_d),$$

où t_1, t_2, \dots, t_d ont comme valeur 1 ou 2, les probabilités conjointes des variables aléatoires Y_1, Y_2, \dots, Y_d . Dans le modèle logistiqua multidimensionnel, les probabilités conjointes de Y_1, Y_2, \dots, Y_d sont paramétrées sous forme de logits marginaux, de risques relatifs logarithmiques marginaux et de contrastes de risques relatifs logarithmiques marginaux. Ce paramétrage peut être exprimé de la manière suivante: $\eta = C^T \log(L\pi)$, où π est le vecteur de dimension $q = 2^d$

$$\pi = (\pi_{11\dots 11}, \pi_{11\dots 12}, \pi_{12\dots 21}, \pi_{22\dots 22})^T,$$

et où les matrices L et C sont les produits tensoriels, respectivement, d'un indicateur marginal et de matrices de contraste appropriées. Les matrices L et C , qui dépendent de la durée d de la période d'observation, sont définies dans une suite récurrente, en commençant par $L_0 = C_0 = 1$, avec

$$L_d = \begin{bmatrix} L_{d-1} \otimes 1_2^T & L_{d-1} \otimes \tilde{L}^T \\ L_{d-1} \otimes 1_2^T & L_{d-1} \otimes \tilde{L}^T \end{bmatrix}$$

$$C_d = \begin{bmatrix} C_{d-1} & 0 \\ 0 & C_{d-1} \otimes \tilde{C} \end{bmatrix},$$

et

a une caractéristique, désignée par le terme «reproductibilité», qui fait que le modèle logistique multidimensionnel s'applique à tout sous-ensemble du vecteur de réponse. Cette caractéristique assure l'égalité des interprétations des paramètres, quel que soit le nombre de variables de réponse, et que des paramètres d'un ordre supérieur soient inclus ou non. Cette propriété rend la régression logistique multi-

dimensionnelle particulièrement intéressante pour l'analyse de données longitudinales, où les observations répétées d'un résultat ont lieu sur un pied d'égalité et où le nombre de ces observations répétées peut varier d'un individu à l'autre. La reproductibilité est également l'élément clé qui permet d'utiliser le modèle pour traiter des réponses incomplètes. Cependant, il est à noter qu'il faut présupposer que l'absence complète de données est aléatoire, si les mêmes paramètres doivent être utilisés pour modéliser aussi bien les réponses complètes que les réponses incomplètes. Les estimations de paramètres sont déterminées selon la méthode du maximum de vraisemblance. Une opération clé est l'inversion de la transformation logistique multidimensionnelle. Dans le cas de plus de trois réponses, il se peut que cette opération ne puisse pas toujours être effectuée, car dans un tel cas il existe des contraintes qui touchent les paramètres (Glonck et McCullagh 1995, Liang, Zeger et Qaish 1992). En outre, la présence de cellules vides peut limiter l'ordre des paramètres qui peuvent être ajustés.

Le modèle log-linéaire est largement utilisé pour modéliser des données binaires multidimensionnelles. Dans le modèle log-linéaire saturé (voir, par ex., Liang et coll. 1992), le paramètre canonique associé à un sous-ensemble des variables a une interprétation sous forme de probabilités conditionnelles, compte tenu des autres variables; par exemple, les paramètres de premier et deuxième ordre sont des logits et des risques relatifs logarithmiques qui dépendent de toutes les autres réponses. Il s'ensuit que le modèle log-linéaire n'est pas repro-

ductible, ce qui en fait un choix moins intéressant que la régression logistique multidimensionnelle pour l'analyse des données longitudinales. Il est néanmoins possible d'élaborer des modèles log-linéaires qui, à l'instar du modèle logistique multidimensionnel, comportent des logits marginaux comme paramètres. Cela mène aux modèles marginaux (Diggle et coll. 1994, chap. 8). Dans ces modèles, la dépendance des probabilités marginales à l'égard de variables explicatives est modélisée de manière distincte de la corrélation existant au sein des unités. Avec cette approche, les paramètres ne sont pas estimés par la méthode du maximum de vraisemblance; on spécifie plutôt uniquement la structure de la corrélation existant entre les observations répétées d'un résultat, et les paramètres sont estimés par la résolution d'équations d'estimation généralisées (EBG), un analogue multidimensionnel de la quasi-vraisemblance (McCullagh et Nelder 1989). Un certain nombre de spécifications de la structure de corrélation ont été proposées; par exemple, Liang et coll. (1992) ont utilisé les risques relatifs marginaux comme

1991 et qui ont été remplacées en 1994, la période d'observation, (notée (1)234, va de 1991 à 1994. Nous utilisons la notation (1)234 pour indiquer le fait que nous n'avons pas eu recours aux observations qui ont été effectuées en 1991.

Tableau 1
Structure des données, taille des échantillons longitudinal et transversal, canton de Vaud, 1992-1995

Première année dans l'échantillon	Époques d'observation des diverses parties de l'échantillon										Période d'observation			
91	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	527	481	612	92	93	94	95	877	4 981
	92	92	93	94	(1)234	52								

La situation d'activité est une variable nominale qui comprend les trois catégories suivantes: «personne occupée», «sans emploi» et «hors de la population active». Dans les exemples que nous donnons dans les sections 4 et 5, nous utilisons une variable binaire qui prend la valeur 1 si la personne sondée occupe un emploi, et la valeur 2 si elle est sans emploi ou hors de la population active. Nous procédons de cette façon uniquement pour simplifier la présentation des modèles logistiques multidimensionnels. Étant donné que la méthode que nous utilisons permet de traiter un nombre arbitraire de catégories, il serait préférable de ne pas regrouper les situations d'activité dans le cas d'une analyse réelle. Dans le cas où il serait néanmoins nécessaire de regrouper certaines situations d'activité, il faut faire preuve de prudence, étant donné que l'hétérogénéité des situations peut introduire des biais.

3. MODÈLES LOGISTIQUES MULTIDIMENSIONNELS

Le modèle logistique multidimensionnel qui a été introduit par Glonek et McCullagh (1995) permet de traiter des réponses multidimensionnelles de type nominal ou ordinal ainsi que des variables explicatives discrètes ou continues. Nous allons considérer ici uniquement des réponses binaires multidimensionnelles et des variables explicatives discrètes. Le modèle logistique multidimensionnel est un exemple de modèle linéaire généralisé (voir McCullagh et Nelder 1989). Sa fonction de liaison, appelée également transformation logistique multidimensionnelle, exprime la distribution conjointe des profils de réponse sous forme de moments marginaux par ordre croissant, les deux premiers moments étant des logits marginaux et des risques relatifs logarithmiques marginaux. La fonction de liaison

précédé l'interview peuvent être reconstitués. Cependant, étant donné que cette reconstitution est fondée sur l'auto-évaluation des répondants, il peut y avoir des imprécisions en ce qui a trait aux situations antérieures et aux moments où les changements ont eu lieu. On peut trouver une analyse des données de l'ESPA fondée sur cette méthode dans Gerfin (1996).

Le présent article est organisé de la manière décrite ci-après. Dans la section 2, nous décrivons les données, un sous-ensemble d'environ 5000 individus tiré de l'ESPA, qui sont utilisées dans les exemples présentés dans les sections 4 et 5. Dans la section 3, nous présentons la régression logistique multidimensionnelle en la comparant avec les modèles log-linéaires et marginaux. Dans la section 4, nous montrons comment on peut utiliser la régression logistique multidimensionnelle pour représenter la structure de dépendance complexe des données de l'ESPA au moyen d'un faible nombre de paramètres. Dans la section 5, nous considérons l'estimation de la variation en comparant la régression logistique multidimensionnelle avec un simple estimateur de variation. Nous montrons que l'utilisation de la régression logistique multidimensionnelle aboutit à un gain d'efficacité. Enfin, dans la section 6, nous présentons nos conclusions et des orientations pour des études ultérieures.

2. DONNÉES DE L'ENQUÊTE SUISSE SUR LA POPULATION ACTIVE

On peut trouver une description détaillée du plan d'échantillonnage et de la méthode de pondération de l'ESPA dans Hülliger, Ries, Comment et Bender (1997). Nous rappelons ici uniquement certains aspects importants de cette enquête. L'ESPA recueille des renseignements sur la situation d'activité de résidents de la Suisse âgés de quinze ans et plus. Depuis le deuxième trimestre de 1991, un échantillon d'environ 16 000 personnes fait l'objet d'interviews annuelles. Cette enquête est fondée sur un plan d'échantillonnage à renouvellement de panel, avec une durée de présence dans l'échantillon de cinq ans. Durant la phase de démarrage, c'est-à-dire entre 1992 et 1996, environ un cinquième de l'échantillon d'origine a été remplacé par un échantillon de renouvellement chaque année. Les unités de l'échantillon de renouvellement vont rester dans le panel pour une période complète de cinq ans.

Dans les exemples que nous donnons dans les sections 4 et 5, nous utilisons les observations relatives à la situation d'activité pour les années 1992 à 1995 qui ont été obtenues auprès des personnes faisant partie de l'échantillon du canton de Vaud. Le tableau 1 indique la structure des données ainsi que la taille des échantillons longitudinal et transversal. En raison du plan d'échantillonnage qui a été choisi, certains des profils de réponse sont incomplets. Par exemple, dans le cas des personnes qui ont été choisies en

Les modèles log-linéaires et les modèles marginaux sont ayant trait à des moments marginaux.

sur les estimations de paramètres de la régression logistique certainement souhaitable d'étudier l'incidence de ces biais sur les estimations de paramètres de la régression logistique multimensionnelle, qui comportent des interprétations

auraient pu être incluses également dans l'analyse. Enfin, il est bien connu que des erreurs de classification peuvent introduire des biais importants dans les probabilités relatives aux profils de réponse observées (voir, par exemple, Pfeiffermann, Skinner et Keith 1998). Il serait certainement souhaitable d'étudier l'incidence de ces biais sur les estimations de paramètres de la régression logistique multimensionnelle, nous y revenons dans la section 3. Nous présentons brièvement ci-après les modèles de transition, les modèles d'effets aléatoires et l'analyse de survie, dans le contexte de l'ESPA. Dans le cas d'un modèle de transition (voir, par ex., Diggle, Liang et Zeger 1994, chap. 10, ou Zeger et Liang 1992), les observations répétées de la situation d'activité sont corrélées parce que les situations d'activité passées ont une incidence sur la situation d'activité actuelle. Le point d'intérêt réside dans les probabilités de transition entre les diverses situations d'emploi; par exemple, la probabilité d'occuper un emploi peut dépendre du fait d'avoir été sans emploi dans le passé. Dans le contexte de la régression, les réponses antérieures sont traitées comme des variables explicatives supplémentaires. Un aspect important est la détermination du nombre de réponses antérieures à inclure comme variables prédictives. Si le modèle relatif aux probabilités de transition est correctement spécifié, on peut traiter des transitions répétées relatives à une personne en particulier comme des événements indépendants et utiliser des méthodes statistiques courantes, comme la régression logistique. Dans le cadre d'un modèle d'effets aléatoires (voir, par ex., Diggle et coll. 1994, chap. 9), la probabilité d'être dans une situation d'activité donnée est une fonction de variables explicatives et les coefficients de régression varient d'un individu à l'autre. Cette variabilité des coefficients de régression reflète l'hétérogénéité naturelle des individus qui est due à des facteurs non mesurés. Compte tenu des coefficients de régression, on présuppose que les observations répétées de la situation d'activité sont indépendantes. La corrélation entre les observations répétées survient uniquement parce que nous ne sommes pas en mesure d'observer les coefficients de régression réels. Cette façon de procéder présente la plus grande utilité lorsque le point d'intérêt réside dans l'inférence relative à des individus plutôt que dans des moyennes de population. Dans le cas de l'analyse de survie, que l'on désigne également par le terme «analyse des antécédents» dans les textes qui traitent d'économétrie (Lancaster 1990), le point d'intérêt est la modélisation des transitions entre les diverses situations d'activité au fil du temps, considérées comme une fonction de variables explicatives. Dans ce cas, l'ESPA, la situation d'activité est observée une fois par an. Les changements qui ont eu lieu durant l'année qui a

Analyse longitudinale de données de l'enquête suisse sur la population active par régression multidimensionnelle

PAUL-ANDRÉ SALAMIN¹

RÉSUMÉ

Dans le cadre d'enquêtes longitudinales, de simples estimations de variations de pourcentages, comme des différences de pourcentages, peuvent ne pas être toujours suffisamment efficaces pour déceler des variations ayant de l'importance sur le plan pratique, notamment dans le cas de sous-populations. Le recours à des modèles, qui peuvent représenter la structure de dépendance de l'enquête longitudinale, peut aider à résoudre ce problème. Une des principales caractéristiques observées dans le cadre de l'enquête suisse sur la population active (ESPA) est la situation d'activité. Étant donné que cette enquête comporte un plan d'échantillonnage à renouvellement de panel, les données qui en sont tirées sont des données nominales multidimensionnelles, et une grande partie des profils de réponse est absente conformément à ce plan. Le modèle logistique multidimensionnel, qui a été introduit par Glonek et McCullagh (1995) comme généralisation de la régression logistique présente de l'intérêt dans ce contexte parce qu'il autorise des observations répétées dépendantes et des profils de réponse incomplets. Nous montrons qu'en utilisant la régression logistique multidimensionnelle, on peut représenter la structure de dépendance complexe de l'ESPA à l'aide d'un petit nombre de paramètres et obtenir des estimations plus efficaces d'une variation.

MOTS CLÉS: Données binaires longitudinales; modèle logistique multidimensionnel; enquête sur la population active.

1. INTRODUCTION

Un des principaux objectifs de l'Enquête suisse sur la population active (ESPA) est la production d'estimations des variations des pourcentages de population relatifs à diverses situations d'activité. Normalement, de simples estimations d'une variation, comme la différence entre les pourcentages observés d'une année à une autre au sujet des personnes qui occupent un emploi, sont calculées pour l'ensemble de la population, ainsi que pour un nombre important de sous-populations. En général, ce résultat est insatisfaisant, car les estimations relatives aux sous-populations peuvent ne pas être toujours suffisamment efficaces pour déceler des variations qui ont une importance sur le plan pratique. Le travail que nous présentons ici répond à la question de savoir si le recours à des modèles qui permettent de représenter la structure de dépendance de l'enquête précitée pouvait contribuer à résoudre ce problème.

Étant donné que l'ESPA est fondée sur un plan d'échantillonnage à renouvellement de panel, nous avons affaire à des données longitudinales nominales dont une grande partie des profils de réponse sont laissés incomplets conformément au plan. Le point central de notre étude est la modélisation des probabilités marginales, et notamment des probabilités d'être dans une situation d'activité donnée, en tant que fonction du temps et d'autres covariables qui définissent des sous-populations. Si les observations répétées de la situation d'activité étaient indépendantes, une façon normale de procéder consisterait à utiliser la régression logistique. Le modèle logistique multidimen-

De nombreuses questions importantes ne sont pas abordées dans le présent article. Parce que les données de l'ESPA sont tirées d'une enquête complexe, on peut soutenir que toute analyse devrait prendre en considération les poids d'échantillonnage (Pfeffermann 1993). Ici, nous utilisons uniquement les données non pondérées. Cependant, on peut montrer, à l'aide de la méthode de pseudo-vraisemblance de Binder (1983), que la régression logistique multidimensionnelle peut être élargie au cas en question (Salamin 1998). La non-réponse est toujours une préoccupation importante dans le cas des enquêtes sur échantillon. Ici, nous considérons uniquement les profils de réponse incomplets qui résultent du renouvellement du panel. Dans ce cas, l'hypothèse de valeurs manquantes complètement aléatoire est raisonnable. Il est à noter, cependant, que la régression logistique multidimensionnelle présente suffisamment de souplesse pour pouvoir intégrer des paramètres supplémentaires en considération des profils incomplets résultant de la perte de personnes faisant partie du panel. Ainsi, les personnes qui ont quitté le panel

sionnel introduit par Glonek et McCullagh (1995) comme généralisation de la régression logistique est intéressant dans ce contexte car il autorise des observations répétées dépendantes et des profils de réponse incomplets. Le but du présent article est de montrer que le recours à la régression logistique multidimensionnelle pour modéliser la structure de dépendance complexe de l'ESPA permet d'obtenir des estimateurs de variation plus efficaces. Bien que nous présentions la méthode en l'appliquant à des données de l'ESPA, il est clair qu'elle peut être appliquée à un éventail de données plus large.

¹ Paul-André Salamin, Statistical Methods Unit, Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel, Switzerland.

REMERCIEMENTS

Le travail de Paul Clarke dans le présent article était financé grâce à une bourse d'études du Conseil de la recherche économique et sociale (prix n° R00429614273); le travail de Ray Chambers était financé grâce au contrat conclu entre l'Office for National Statistics et l'Université de Southampton pour la prestation de services de recherches en méthodologie statistique. Les deux auteurs aimeraient remercier les examinateurs, dont les observations et les conseils pratiques ont aidé à rendre la version définitive du manuscrit considérablement plus compréhensible.

BIBLIOGRAPHIE

LITTLE, R.J.A. (1982). Models for nonresponse in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

LITTLE, R.J.A. (1985). Nonresponse adjustments in longitudinal surveys: models for categorical data. *Bulletin de l'Institut International de Statistique*, 15, 1-15.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

STASNY, E.A. (1986). Estimating gross flows using panel data with nonresponse: an example from the Canadian Labour Survey. *Journal of the American Statistical Association*, 81, 42-47.

STASNY, E.A., et FIENBERG, S.E. (1985). Some stochastic models for estimating gross flows in the presence of nonrandom nonresponse. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 25-39.

VANSKI, J.E. (1985). Uses of gross change data in assessing demographic labor market dynamics. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 9-12.

FTZMAURICE, G.M., LAIRD, N.M., et ZAHNER, G.E.P. (1996). Multivariate logistic models for incomplete binary responses. *Journal of the American Statistical Association*, 91, 99-107.

HOGUE, C.R. (1985). History of the problems encountered in estimating gross flows. *Proceedings of the Conference on Gross Flows in Labor Force Statistics*, U.S. Bureau of the Census, 1-8.

la population active des membres du ménage. La non-réponse dans les enquêtes sur les ménages peut survenir pour plusieurs raisons, par ex., un refus, un non-contact, un déménagement ou un renouvellement de l'échantillon. Le modèle actuel peut facilement être étendu pour modéliser des tendances plus complexes de non-réponse en précisant l'indicateur de non-réponse comme étant une variable polynôme et en paramétrant le modèle de non-réponse conformément aux tendances complexes de non-réponse. Il est également à remarquer que nous ne supposons pas que le modèle au niveau du ménage est une représentation exacte du comportement de non-réponse du ménage; plutôt, nous supposons que le modèle au niveau de la personne offre une approximation de la dynamique de non-réponse au sein du ménage.

Un problème important, mis en évidence par les résultats provenant de l'étude en simulation, est notre hypothèse que le comportement des flux de la population active individuels est homogène au sein des ménages. De toute évidence, il s'agit d'une hypothèse qui n'est pas réaliste. Le modèle est facilement étendu en précisant les flux de la population active et les probabilités de flux de non-réponse comme modèles de régression pour tenir compte des renseignements liés à la covariable au niveau de la personne, au niveau du ménage ou à un niveau plus élevé. Par exemple, les probabilités de flux de la population active pourraient être précisées comme étant une régression multinomiale-logistique:

$$\log \left(\frac{\omega_{hi}(ab)}{\omega_{hi}(11)} \right) = \beta_{(ab)}^0 + \beta_{(ab)}^1 \mathbf{x}_T^{h_{hi}}$$

où $\omega_{hi}(ab)$ désigne la probabilité d'une personne i dans un ménage h du flux de la population active (a, b), $\mathbf{x}_T^{h_{hi}}$ est un vecteur de covariable (rangée), et $(\beta_{(ab)}^0, \beta_{(ab)}^1)$ sont les coefficients de régression pour le multinomial-logit (a, b). Cependant, l'ajustement de ces modèles nécessite que l'on émette des hypothèses d'indépendance conditionnelles au sujet de la relation entre les distributions des covariables, les flux de la population active et les flux de non-réponse parce que les renseignements sur les covariables peuvent être manquants dans le cas des ménages qui ne répondent pas. Une autre solution est de permettre des covariables hétérogènes entre les flux de la population active du ménage, à l'aide d'effets aléatoires, en adoptant des hypothèses relativement à la distribution des différences entre les ménages. L'ajustement de ces modèles est également compliqué et nécessiterait, par exemple, une méthode de Monte Carlo de chaîne de Markov pour effectuer l'intégration nécessaire. Si S n'est pas un échantillon aléatoire simple, on peut incorporer des variables auxiliaires du plan dans le processus d'ajustement en utilisant le cadre de régression que nous venons de décrire.

compte. Pour produire la non-réponse au niveau du ménage dont il faut tenir compte, il est nécessaire de prévenir $n_h(ab) - n_h\omega(ab)$ en étendant (1) de façon à permettre des flux de la population active différentiels entre les ménages. Ces extensions du modèle des flux de la population active font l'objet de la section 4.

La figure 1b) illustre deux résultats anormaux qui contredisent l'explication ci-dessus lorsque N_B est le modèle véritable. Tout d'abord, le biais de l'estimation du modèle N_B au niveau de la personne augmente à mesure que n_h augmente. Cependant, d'autres simulations avec un ménage d'une taille $n_h = 10$ ont révélé que le biais de l'estimation au niveau de la personne est zéro. Ainsi, la non-réponse asymptotique dont on n'a pas à tenir compte est également évidente lorsque N_B est vrai, mais n_h doit être importante avant que son incidence devienne apparente pour le modèle N_B . Deuxièmement, le biais des estimations du modèle dont on n'a pas à tenir compte au niveau de la personne est faible, presque zéro, lorsque N_B est vrai. Ce faible biais diminue encore plus à mesure que n_h augmente, conformément à l'ignorabilité asymptotique, mais nous devons encore parvenir à une explication satisfaisante quant à savoir pourquoi les modèles dont on peut ne pas tenir compte se comparent si bien dans cette situation. Des études plus approfondies sont nécessaires pour examiner cette constatation.

4. DISCUSSION

Dans les sections 3 et 4, on a démontré grâce à une étude en simulation que la modélisation de la non-réponse dont il faut tenir compte au niveau du ménage lors de l'estimation des flux bruts de la population active provenant d'enquêtes sur les ménages entraîne une réduction du biais dans les estimations des flux, par rapport aux flux provenant des modèles au niveau de la personne. S'il s'agit d'un modèle de non-réponse dont on n'a pas à tenir compte, il n'est pas nécessaire de recourir à des modèles au niveau du ménage parce que les modèles au niveau du ménage et au niveau de la personne sont équivalents. En outre, on a démontré que le contrôle de la non-réponse au niveau du ménage n'élimine pas nécessairement tout le biais provenant des estimations des flux de la population active. La spécification correcte du modèle de non-réponse est toujours considérée impérative, quoique le fait de tenir compte de la structure des données du ménage peut entraîner un peaufinage des estimations des flux si le modèle de non-réponse est mal spécifié. En particulier, nous démontrons que les estimations au niveau du ménage sont moins biaisées que leurs estimations équivalentes au niveau de la personne.

Notre modèle de non-réponse est une extension de l'idée qui veut que la non-réponse peut dépendre des caractéristiques d'une unité, dans le présent cas les flux de

3. ETUDE EN SIMULATION

3.1 Procédure de simulation

Nous avons eu recours à une étude en simulation pour examiner les conséquences de ne pas tenir compte de la structure du ménage dans le cas des données d'enquêtes sur les ménages pour comparer les estimations des flux bruts de la population active pour les modèles au niveau de la personne et au niveau du ménage. À cette fin, on a produit des données d'enquêtes sur les ménages à l'aide de l'échantillonnage de Monte Carlo. Chaque ensemble de données de l'échantillon se composait de 10 000 personnes réparties dans des ménages de taille $n_h = k$ pour la totalité des h . Au sein de chaque ménage, on a produit les flux de la population active à partir de (1) et on a produit le flux de non-réponse à partir de (2) en vertu d'un des modèles N_A ou N_B . On a rendu les données incomplètes en regroupant chaque tableau des flux de la population active des données complètes pour qu'il soit uniforme avec le flux de non-réponse du ménage. Au total, 1 000 ensembles de données indépendantes ont été produits de cette façon.

Les paramètres de la population active pour produire les flux de la population active sont illustrés dans le tableau suivant:

$\omega(ab)$			
	1	2	3
a	0.43	0.02	0.015
	0.245	0.16	0.035
	0.035	0.01	0.05
b	1	2	3

Il s'agit de toute évidence d'une population en récession étant donné que la probabilité de passer de personne employée à personne sans emploi est très grande ($\omega(12) = 0.245$). En vertu des modèles N_A et N_B les paramètres de la population sont

$\lambda(i)$	$\theta(i)$			
		1	2	3
!	0.2	0.5	0.2	0.8
	0.8	0.5	0.2	0.8
	0.5	0.5	0.2	0.8

On doit remarquer que ces valeurs des paramètres ne représentent pas un comportement des flux de non-réponse réaliste, elles ont été choisies dans le but d'illustrer la présente méthodologie. Cependant, cela n'influe pas sur les conclusions générales de l'article, qui sont également

3.2 Résultats de la simulation

pertinentes pour des valeurs réalistes des véritables probabilités de non-réponse.

On obtient les estimations pour les modèles au niveau de la personne en corrigeant (4) avec $n_h = 1$ pour chaque ensemble de données incomplètes. La figure 1 résume les distributions d'échantillonnage de l'estimation de la vraisemblance maximale au niveau de la personne de $\omega(12)$, pour les modèles de non-réponse I_A , I_B , N_A et N_B (les estimations pour les modèles dont on n'a pas à tenir compte I_A et I_B sont incluses parce que les deux donnent les mêmes estimations des flux de la population active). Les traits verticaux représentent les intervalles entre le percentile 2,5 et le percentile 97,5 de chaque distribution de l'échantillonnage de l'estimation, et les points en caractères gras représentent ses médianes. On obtient trois distributions pour chaque estimation au niveau de la personne: la distribution la plus à gauche survient lorsque la taille du ménage est $k = 1$, les données simulées n'ont aucune structure de ménage; et si on lit de gauche à droite, les deux distributions suivantes sont celles que l'on obtient lorsque la taille du ménage est $k = 2$ et $k = 5$, respectivement. Le trait vertical plein désigne la véritable probabilité de flux, $\omega(12) = 0.0245$. Le comportement de la distribution d'échantillonnage de $\hat{\omega}(12)$ dans la présente étude reflète celui des autres estimations des flux bruts de la population active.

La figure 1a) en résume les distributions d'échantillon-nage lorsque N_A est le véritable modèle. Si le modèle ajusté au niveau de la personne est I_A , I_B ou N_B , les estimations des flux bruts de la population active ont d'importants biais, quelle que soit la taille du ménage. Comme on devrait s'y attendre, l'estimation de la médiane pour le modèle correct N_A n'est pas biaisée si $k = 1$ et a un petit biais pour $k = 2$ et $k = 5$ (quoique ce biais soit plus faible dans le cas de $k = 5$ que dans celui de $k = 2$). La diminution du biais lorsque la valeur de k augmente est également apparente dans le cas des estimations au niveau de la personne I_A , I_B et N_B . Ce comportement n'est pas prévu étant donné qu'il semble naturel de s'attendre à ce que le biais des estimations au niveau de la personne augmente avec la taille du ménage. Les résultats sont légèrement différents dans la figure 1b) lorsque N_B est vrai. Dans le présent cas, l'estimation pour le modèle N_B au niveau de la personne devient plus biaisée à mesure que la valeur de k augmente, mais le biais diminue pour les modèles mal spécifiés I_A , I_B et N_B au niveau de la personne. En outre, les estimations mal spécifiées pour I_A et I_B comportent un petit biais lorsqu'on les compare à celles du modèle mal spécifié N_A .

On traite de ces résultats à la section 3.3.

les paramètres de non-réponse sont limités de sorte que

- Modèles dont on n'a pas à tenir compte.
- Modèle I_A^A : Probabilité de non-réponse constante,

$$\psi(nv | ab) = \lambda_{1-n}(1 - \lambda)_n \times \lambda_{1-v}(1 - \lambda)_v,$$
 qui a 1 paramètre, λ , la probabilité d'une personne qui ne répond pas;
- Modèle I_B^B : Indépendant de la situation vis-à-vis de la population active, mais probabilités de non-réponse différentes à t_1 et t_2 ,

Lorsque $n_h > 1$, déterminer l'estimabilité des paramètres est plus difficile parce que (4) a une expression en forme

(1996) utilise une méthode numérique pour déterminer l'estimabilité qui implique une démonstration que la matrice d'information est régulière dans le voisinage de l'estimation de vraisemblance maximale. Cependant, ce n'est non seulement peu pratique pour les problèmes d'une dimension élevée, mais évaluer la matrice d'information par la méthode de Fisher est souvent coûteux.

difficile dans ce cas. Au lieu, nous adoptons une approche

pragmatique pour déterminer l'estimabilité des paramètres: tout d'abord, nous restreignons l'attention aux modèles qui remplissent la condition nécessaire pour l'estimabilité lorsque $u = 1$; et deuxièmement, on utilise différentes valeurs de départ pour chaque ajustement. Si les valeurs de départ différentes révèlent une estimation de vraisemblance

maximale non unique, ou si les estimations des paramètres demeurent inchangées par rapport à leurs valeurs de départ, alors on suppose que les paramètres du modèle ne peuvent être estimés.

2.4 Modèles de non-réponse

Pour permettre d'obtenir les estimations des paramètres θ et ψ doivent être limitées à partir des données observées,

conformément aux hypothèses au sujet du mécanisme de non-réponse. Les paramètres de non-réponse sont interprétés, comme des probabilités de non-réponse individuelles, mais en ce qui concerne le cadre de ménages établis jusqu'à maintenant, il n'est pas approprié de parler de non-réponse qui ne répond pas. Cependant, on peut dire que

personnes qui ne répondent pas. Cependant, en raison ce sont les personnes au sein des ménages qui déterminent un flux de non-réponse d'un ménage, et non le ménage lui-même. Par conséquent, les contraintes sont placées sur les paramètres de non-réponse au niveau de la personne qui s'applique au niveau du ménage par la dépendance fonctionnelle de $\pi(uv | u^h)$ sur ψ dans (2). Par exemple, si

Le modèle des flux de la population active est multinomial avec une fonction de probabilité

$$(1) \quad \Pr(N_h = n_h; \omega) = n_h! \prod_{a,b} \frac{\omega(n_h(ab)!)^{\omega(ab)}}{\omega(ab)^{n_h(ab)}}$$

Représentons les données observées par $\{n_h^*\}$. Tel qu'on l'a indiqué à la section 2.1, les données observées pour les ménages qui répondent à t_1 et t_2 sont la classification croisée complète du tableau 1, à savoir $n_h^* = n_h$. De même, si $h \in S_{10}$ alors $n_h^* = (n_h^*(1+), n_h^*(2+), n_h^*(3+))$; si $h \in S_{01}$ alors $n_h^* = (n_h^*(+1), n_h^*(+2), n_h^*(+3))$; et si $h \in S_{00}$ alors $n_h^* = n_h$. On obtient la contribution du ménage $h \in S_{uv}$ à la vraisemblance des données observées en faisant le total de $L_h(\omega, \psi; n_h^*, (u, v))$ pour toutes les valeurs possibles que la classification croisée complète 3×3 des flux de la population active peut prendre compte tenu de la marge observée. Représentant cet ensemble de tableaux par $n_h : n_h^*$, la vraisemblance des données observées pour S est

$$(4) \quad L(\omega, \psi; \{n_h^*, r_h^*\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; n_h^*, (u, v)).$$

L'ajustement du modèle nécessite le calcul de (4) à chaque étape d'un processus d'optimisation itérative. Sur le plan des calculs, ceci est exigeant parce que la fonction de vraisemblance des données complètes doit être additionnée explicitement pour les données manquantes. Par exemple, les données observées pour $h \in S_{10}$ est $n_h^* = (n_h^*(1+), n_h^*(2+), n_h^*(3+))$ et la contribution de vraisemblance de ce ménage à la vraisemblance des données observées est

$$\sum_{n_h: n_h^*} L_h(\omega, \psi; n_h, (1, 0)).$$

Pour calculer explicitement cette contribution, chaque tableau des données complètes 3×3 n_h pour n_h^* fixe est produit et $L_h(\omega, \psi; n_h, (1, 0))$ est évaluée pour chacun. Dans le cas d'un ménage de taille $n_h = 5$, il y a au moins 21 tableaux possibles et au plus 108 tableaux possibles, selon les valeurs de la marge fixe; lorsque $n_h = 15$, un ménage d'une très grande taille, les nombres respectifs sont de 136 et de 9261. Une procédure semblable est utilisée pour $h \in S_{01}$, sauf qu'ici $n_h^* = (n_h^*(+1), n_h^*(+2), n_h^*(+3))$ est la marge fixe. Si $h \in S_{00}$ alors aucune donnée n'est observée au sujet de la situation vis-à-vis de la population active, seulement la taille du ménage n_h . Donc chaque tableau 3×3 ayant le total n_h doit être produit et la fonction de vraisemblance être calculée pour chacun: lorsque $n_h = 5$ il y a 1287 tableaux et lorsque $n_h = 15$ il y a 490314 tableaux. Il est impossible, en ce qui concerne le temps d'exécution machine, de calculer ces sommes directement. Le nombre de calculs explicites peut être réduit si l'on reconnaît que chaque ménage est défini uniquement par ses fréquences des flux de la population active observés et son flux de non-réponse. Ainsi, la somme des données manquantes ne doit être effectuée qu'une fois pour un ménage qui a des fréquences de flux de la population active et un flux de non-réponse donnés; la contribution de ce ménage à la vraisemblance est alors élevée à la puissance du nombre de ménages définis de façon semblable dans S .

La probabilité pour le ménage h d'avoir un flux de non-réponse (u, v) est

$$\pi(uv | n_h) = \Pr(R_h = (u, v) | N_h = n_h; \psi) = \frac{1}{\sum_{a,b} n_h(ab) \psi(uv | ab)},$$

lorsque $u, v = 0, 1$, à savoir une moyenne pondérée des paramètres du modèle de non-réponse. En établissant $n_h = 1$, on peut constater que $\psi(uv | ab) > 0$ est la probabilité d'un ménage d'une taille un (c'est-à-dire une personne) qui a un flux de non-réponse (u, v) si l'a un flux de population active (a, b) . Ainsi, $\sum_{u,v} \psi(uv | ab) = 1$ et $\psi = (\psi(11 | 11), \psi(01 | 11), \psi(00 | 33))$ est le vecteur des paramètres de non-réponse, dont 27 sont libres. Avant de définir la fonction de vraisemblance des données complètes, partageons S en 4 sous-ensembles exhaustifs et mutuellement exclusifs

$$S = S_{11} \cup S_{01} \cup S_{10} \cup S_{00},$$

où $S_{uv} = \{h : r_h = (u, v)\}$ est le sous-ensemble des ménages ayant un flux de non-réponse (u, v) . Ainsi, étant donné que S est un échantillon aléatoire simple des ménages, la fonction de vraisemblance des données complètes est

$$(3) \quad L(\omega, \psi; \{n_h^*, r_h^*\}) = \prod_{u,v} \prod_{h \in S_{uv}} L_h(\omega, \psi; n_h^*, (u, v)),$$

où $L_h(\omega, \psi; n_h^*, (u, v))$ est la contribution du ménage $h \in S_{uv}$ à la vraisemblance, le produit de (1) et (2).

2.3 Ajustement des modèles

2.3.1 Estimation de vraisemblance maximale

Étant donné que les données complètes ne sont pas disponibles, (3) doit être modifié de façon à donner la

réponse dont il faut tenir compte. Nous simulons alors des données d'enquêtes sur les ménages à l'aide de ces modèles au niveau du ménage afin de démontrer l'utilité éventuelle de notre approche; premièrement, on démontre que les estimations des flux bruts de la population active au niveau de la personne sont biaisées lorsqu'elles sont ajustées en fonction des données d'enquêtes sur les ménages; deuxièmement, on compare le biais des estimations des flux bruts au niveau de la personne à celui au niveau du ménage afin de démontrer les avantages d'ajuster les modèles au niveau des ménages aux données d'enquêtes sur les ménages. Enfin, nous résumons les conclusions de nos études en simulation et nous présentons des orientations pour des études ultérieures.

2. UN MODÈLE POUR LA NON-RÉPONSE AU NIVEAU DU MÉNAGE

2.1 Les données

Un flux brut est la probabilité ou la fréquence de personnes au sein de la population qui font une transition d'états entre deux points dans le temps, t_1 et t_2 ($t_1 < t_2$). Les flux bruts de la population active renvoient aux transitions entre les trois grands états de la population active: 1 = «personne occupée», 2 = «personne sans emploi» et 3 = «hors de la population active», la dernière catégorie faisant référence aux personnes économiquement inactives telles les personnes à la retraite et les étudiants. Soit S qui représente un échantillon aléatoire simple de ménages, indexé par h . Au sein du ménage h , il y a n_h personnes admissibles, dont $n_h(ab)$ ont un flux de population active (a, b) entre t_1 et t_2 , où $\sum_{a,b} n_h(ab) = n_h$ et $a, b = 1, 2, 3$. Nous disons que $\{n_h(ab)\}$ sont les données complètes, c'est-à-dire les fréquences qui seraient observées en l'absence de non-réponse.

Le tableau 1 illustre les données de flux de la population active complètes pour le ménage h comme étant un tableau de contingence de 3×3 . Si h répond les deux fois, les données observées sont les cellules de ce tableau à double entrée. Cependant, si le ménage ne répond pas à t_1 ou t_2 , les données observées correspondent aux marges du tableau: $n_h(1+)$, $n_h(2+)$, $n_h(3+)$ sont les données observées si h répond à t_2 mais ne répond pas à t_1 ; et $n_h(+1)$, $n_h(+2)$, $n_h(+3)$ sont les données observées si h répond à t_1 mais ne répond pas à t_2 . (Un indice remplacé par «+» représente la somme de tous les niveaux de cet indice.) En outre, si h ne répond pas à t_1 et t_2 , les données observées sont la taille du ménage, n_h , que nous supposons connues et fixes entre t_1 et t_2 .

Tableau 1
Données complètes des flux de la population active pour le ménage h

État	t_2	t_1			n_h
		1	2	3	
t_1	1	$n_h(11)$	$n_h(12)$	$n_h(13)$	$n_h(1+)$
	2	$n_h(21)$	$n_h(22)$	$n_h(23)$	$n_h(2+)$
	3	$n_h(31)$	$n_h(32)$	$n_h(33)$	$n_h(3+)$
		$n_h(+1)$	$n_h(+2)$	$n_h(+3)$	n_h

2.2 Spécification du modèle

Il ne convient pas de traiter le comportement de non-réponse de personnes au sein des ménages comme étant indépendant dans les enquêtes sur les ménages. Dans la Labour Force Survey, par exemple, un membre admissible du ménage détermine si le ménage peut être interviewé. Par conséquent, s'il n'y a aucune personne admissible que l'on peut contacter, chaque personne du ménage ne répond pas. Pour construire un modèle de la non-réponse au niveau du ménage, nous prenons les idées à l'origine de la non-réponse au niveau de la personne et nous les étendons au ménage en considérant qu'un ménage est une entité ayant son propre flux de non-réponse entre t_1 et t_2 . Pour permettre une non-réponse dont il faut tenir compte, la probabilité d'un flux de non-réponse au sein d'un ménage est modélisée en tant qu'une fonction de ses flux individuels de la population active, que nous devons maintenant décrire. Soit $N_h = (N_h(11), N_h(12), \dots, N_h(33))$ le vecteur aléatoire des fréquences des flux de la population active pour le ménage h , où $N_h(ab)$ est la variable aléatoire dont l'extrait correspond au nombre de personnes du flux de la population active (a, b) , $a, b = 1, 2, 3$. En outre, désignons le vecteur aléatoire pour le flux de non-réponse du ménage h par $R_h = (R_{h1}, R_{h2})$, où

$$R_{hj} = \begin{cases} 1, & \text{si ménage répond à } t_j \\ 0, & \text{sinon} \end{cases}$$

est la variable aléatoire de l'état de non-réponse pour h à t_j , $j = 1, 2$. Les réalisations de ces quantités aléatoires sont désignées par n_h et r_h . Nous supposons maintenant que n_h et r_h sont connus et formons la probabilité conjointe de N_h et R_h comme

$$\Pr(N_h = n_h, R_h = r_h) = \Pr(N_h = n_h) \Pr(R_h = r_h | N_h = n_h),$$

où $\Pr(N_h = n_h)$ est le modèle des flux de la population active et $\Pr(R_h = r_h | N_h = n_h)$ est appelé le modèle des flux de non-réponse.

Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage

PAUL S. CLARKE et RAY L. CHAMBERS¹

RÉSUMÉ

La mesure des flux bruts de la population active est un objectif important des enquêtes continues sur la population active effectuées par un grand nombre d'offices nationaux de la statistique. Cependant, il est bien connu que l'estimation de ces flux peut être compliquée par une non-réponse, des erreurs de mesure, un renouvellement de l'échantillon et des effets complexes du plan de sondage. Le présent article, inspiré par des modèles de non-réponse dans les enquêtes, porte sur l'estimation des flux bruts tout en apportant des ajustements en fonction de la non-réponse dont il faut tenir compte. Les approches antérieures basées sur un modèle en ce qui concerne l'estimation des flux bruts supposaient que la non-réponse était un processus au niveau de la personne. Nous proposons une catégorie de modèles qui permettent une non-réponse dont il faut tenir compte au niveau du ménage. On a recours à une étude en simulation pour démontrer que les estimations des flux bruts de la population active au niveau de la personne provenant des données d'enquêtes sur les ménages peuvent être biaisées et que les estimations en fonction de modèles au niveau du ménage peuvent permettre de réduire ce biais.

MOTS CLÉS : Flux bruts; enquêtes sur les ménages; non-réponse dont il faut tenir compte.

1. INTRODUCTION

On définit de façon générale les flux bruts de la population active comme étant des transitions dans le temps entre les trois grands états de la population active, personnes occupées, personnes sans emploi et personnes économiquement inactives. Les estimations des flux bruts sont un outil important dans l'étude de la dynamique de la population active (par exemple, voir Vanski 1985). Les enquêtes continues à grande échelle telle la Labour Force Survey en Grande-Bretagne et la Current Population Survey aux États-Unis fournissent des données pour l'estimation des flux bruts. Cependant, la non-réponse, l'erreur de mesure, le renouvellement de l'échantillon et les effets complexes du plan de sondage ont une incidence sur l'estimation des flux bruts de ces enquêtes. Hogue (1985) traite de ces facteurs et d'autres qui ont une incidence sur l'estimation des flux bruts. Dans le présent article, nous nous concentrons sur le problème de la non-réponse.

Nous supposons qu'un mécanisme de non-réponse a pour conséquence que les données observées sont incomplètes. Si la probabilité de la non-réponse dépend des données manquantes, alors le mécanisme de non-réponse est un mécanisme dont il faut tenir compte (Rubin 1976). L'approche basée sur des modèles pour analyser des données incomplètes d'enquête est détaillée dans Little (1982). Les approches basées sur des modèles relativement à l'estimation des flux bruts de la population active font intervenir la modélisation des flux de la population active et du mécanisme de non-réponse tout en assurant l'ajustement des deux modèles aux données incomplètes. Des

exemples de ces modèles sont données dans Stasny et Fienberg (1985), Stasny (1986) et, dans le cas de la non-réponse dont il faut tenir compte, dans Little (1985). Nous disons de ces modèles qu'ils sont au niveau de la personne parce que les personnes sont modélisées comme répondant ou ne répondant pas, indépendamment des autres personnes échantillonnées.

Tant la Labour Force Survey que la Current Population Survey sont des exemples d'enquêtes sur les ménages, c'est-à-dire des enquêtes se fondant sur un échantillon aléatoire des ménages plutôt que sur des personnes. Les enquêtes sur les ménages peuvent donner lieu à un comportement corrélé de non-réponse au sein des ménages. Par exemple, dans la Current Population Survey, un seul membre du ménage (habituellement le chef du ménage) agit en tant que représentant des membres du ménage; ainsi, si le membre choisi du ménage ne répond pas, les autres membres du ménage ne répondent pas non plus. Il s'ensuit que, en raison du comportement corrélé de non-réponse au sein du ménage, les modèles de non-réponse au niveau de la personne ne conviennent pas à l'estimation des flux bruts de la population active qui utilise des données d'enquêtes sur les ménages.

Dans le présent article, nous proposons une catégorie de modèles pour les flux de la population active au niveau de la personne et une non-réponse au niveau du ménage qui tient compte du comportement corrélé de non-réponse au sein du ménage. On présente également un certain nombre de modèles plausibles de non-réponse qui sont estimables à partir des données observées, tant pour ce qui est de la non-réponse dont on n'a pas à tenir compte que pour la non-

¹ Paul S. Clarke et Ray L. Chambers, Department of Social Statistics, University of Southampton, Highfield, Southampton, SO17 1BJ, United Kingdom.

- MARTINI, A. (1988). Retrospective versus panel data in estimating labour force gross flows: comparing SIPP and CPS. *Proceedings of the Social Science Section, American Statistical Association*, 109-114.
- MARTINI, A. (1989). Seam effect, recall bias, and the estimation of labour force transition rates from SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 387-392.
- MENG, X.L., et RUBIN, D.B. (1993). Maximum likelihood estimation via ECM algorithm: a general framework. *Biometrika*, 80, 267-278.
- MCCORMICK, M., BUTLER, D., et SINGH, R. (1992). Investigating time-in-sample effects for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 554-559.
- O'MURCHEARTAIGH, C. (1996). Measurement errors in panel surveys: implications for survey design and for survey instruments. *Proceedings of the Scientific Reunion of the Italian Statistical Society*, 1, 207-218. Rimini: Maggioni.
- PFEFFERMAN, D., SKINNER, C.J., et HUMPHREYS, K. (1998). The estimation of gross flows in the presence of measurement error using auxiliary variables. *Journal of the Royal Statistical Society, Series A*, 161, 13-32.
- POTERBA, J.M., et SUMMERS, L.H. (1986). Reporting errors and labor market dynamics. *Econometrica*, 54, 1319-1338.
- SINGH, A.C., et RAO, J.N.K. (1995). On the adjustment of gross flows estimates for classification errors with application to data from the Canadian Labor Force Survey. *Journal of the American Statistical Association*, 90, 1-11.
- SKINNER, C.J., et TORELLI, N. (1993). Measurement error and the estimation of gross flows from longitudinal economic data. *Statistica*, 3, 391-405.
- U.S. DEPARTMENT OF COMMERCE (1991). *SIPP User's Guide*. Washington D.C.
- van de POL, F., et LANGEBEINE, R. (1990). Mixed Markov latent class models. Dans *Sociological Methodology*, (Ed. C.Clogg), 213-247. Oxford: Blackwell.
- van de POL, F., et LANGEBEINE, R. (1992). Analysing Measurement Error in Quasi-experimental Data: An Application of Latent Class Models to Labour Market Data. Document de travail de European Scientific Network on household Panel Studies, 57, Colchester, University of Essex.
- van de POL, F., et LANGEBEINE, R. (1997). Separating change and market data. Dans *Survey Measurement and Process Quality*, (Ed. L. Lyberg et al.), 671-688. New York: Wiley.
- VERMUNT, J.K. (1993). Log-linear and event history analysis with missing data using the EM algorithm. WORK PAPER 93.09.015/7, Tilburg University.
- YOUNG, N. (1989). Wave seam effects in the SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 393-398.

2) Patron des restrictions imposées aux probabilités de réponse

Tableau B.1

Mois d'observation

Probabilité	d'observer un état	Mars 91	Juin 91 & Sept. 91	Déc. 90 & Déc. 91
Obo	étant donnée une transition latente	1	1	1
Cbo		2	2	2
Hbo		3	3	3
Ooc		L	L	L
Coc		L	L	L
Hoc		*	*	*
Ooh		L	L	L
Coh		*	*	*
Hoh		L	L	L
Oco		L	L	L
Cco		L	L	L
Hco		*	*	*
Occ		4	4	4
Ccc		5	5	5
Hcc		6	6	6
Och		*	*	*
Cch		L	L	L
Hch		L	L	L
Oho		L	L	L
Cho		L	L	L
Hho		L	L	L
Ohc		*	*	*
Chc		L	L	L
Hhc		L	L	L
Ohh		7	7	7
Chh		8	8	8
Hhh		9	9	9

Des chiffres égaux indiquent que les probabilités de réponse sont considérées comme étant égales.
* indique une probabilité établie à 0.
L désigne un paramètre libre.

BIBLIOGRAPHIE

ABOWD, J.M., et ZELLNER A. (1985). Estimating gross labour force flows. *Journal of Business and Economics Statistics*, 3, 254-283.

BASSI, F., CROON, M., HAGENAARS, J., et VERMUNT, J. (1995). Estimating Latent Turnover Tables When Data are Affected by Correlated and Uncorrelated Classification Errors. *WORC PAPER 95.12.25/7*, Tilburg University.

BERNARD, H.R., KILLWORTH, P., KRONENFELD, D., et SAILER, L. (1984). The problem of informant accuracy: The validity of retrospective data. *Annual Review of Anthropology*, 13, 495-517.

CHAKRABARTY, R.P., et WILLIAMS, T.R. (1989). Time-in-sample biases in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 309-314.

CHUA, T.C., et FULLER, W.A. (1987). A model for multinomial response error applied to labor flows. *Journal of the American Statistical Association*, 82, 46-51.

CITRO, C.F., et KALTON, G. (1993). *The Future of the SIPP*. Washington, D.C.: National Academy Press.

DUNCAN, G., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.

GOODMAN, L.A. (1973). The analysis of a multidimensional contingency table when some variables are posterior to the others. *Biometrika*, 60, 179-192.

GOODMAN, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 79, 1178-1259.

GOODMAN, L.A. (1981). Three elementary views of log-linear models for the analysis of cross-classifications having ordered categories. Dans *Sociological Methodology*, (Ed. S.Leinhardt), 193-293. San Francisco: Jossey Bass.

HABERMAN, S.J. (1978). *Analysis of Qualitative Data. Vol.1. Introductory Topics*. New York: Academic Press.

HAGENAARS, J.A. (1988). Latent structure models with direct effects between the indicators, local dependence models. *Sociological Methods and Research*, 16, 379-405.

HAGENAARS, J.A. (1990). *Categorical Longitudinal Data: Log-linear, Panel, Trend and Cohort Analysis*. Newbury Park: Sage.

HAGENAARS, J.A. (1997). Categorical Causal Modeling: Directed Loglinear Models With Latent Variables. *WORC PAPER 97.04.002/7*, Tilburg University.

HUBBLE, D.L., et JUDKINS, D.R. (1989). Measuring the Bias in Gross Flows in the Presence of Autocorrelated Response Errors. *SIPP Document de travail no. 8712*, U.S. Bureau of the Census.

HU, S., SMAN, R., MEHRAN, F., et VERMA, M. (1990). *Surveys of Economically Active Population, Employment and Underemployment: an ILO Manual on Concepts and Definitions*. Geneva: ILO.

JORESKEÖG, K.G., et SÖRBOM, D. (1988). *Listrel 7: A Guide to the Program and Applications*. Chicago: SPSS INC.

KALTON, G., et CITRO, C.F. (1993). Enquête par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.

KALTON, G., et MILLER, M.W. (1991). The seam effect with social security income in the SIPP. *Journal of Official Statistics*, 7, 235-245.

LANGHEINE, R., PANNEKOEK, J., et van de POL, F. (1995). Bootstrapping goodness-of-fit measures in categorical data analysis. *Sociological Methods and Research*, 24, 492-516.

LAZARSFELD, P.F., et HENRY, N.W. (1968). *Latent Structure Analysis*. New York: Houghton Mifflin.

MAGNAC, T., et VISSER, M. (1995). Transition Models With Measurement Errors. Document de travail, Institut National de la Recherche Agronomique (INRA), Paris.

MARQUIS, K.H., et MOORE, J.C. (1989). Some response errors in SIPP with thoughts about their effects and remedies. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 381-386.

ANNEXE A

Spécification finale de modèle pour les données de la
SIP, en ce qui a trait aux probabilités conditionnelles
 (1) Décomposition de modèle de base

$$z_g = P(G = g)$$

$$(A1) \quad \pi_{j_1}^1 = P(Y_1 = j_1)$$

$$(A2) \quad \pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3)$$

$$(A3) \quad \pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3)$$

$$(A4) \quad \pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3)$$

$$(A5) \quad \pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3)$$

$$(A6) \quad \pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3)$$

$$(A7) \quad \pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_1 = j_1, Y_2 = j_2, Y_3 = j_3)$$

g varie au cours de 1, 2, 3 et 4; l_t et j_t , $t = 1, 2, 3, 4$, varient dans le cas des catégories O, C et H, m_t , $t = 1, 2, 3, 4$ varient dans le cas des catégories «emploi» et «pas d'emploi».

(2) Contraintes imposées aux probabilités conditionnelles

$$q_{j_1 j_2 j_3}^{j_1 j_2 j_3}$$

$$= P(Y_t = l_{t+1} | Y_t = j_t, Y_{t+1} = l_{t+1}, Y_{t+2} = j_{t+2}, G = g) = 1$$

$$(A8) \quad \text{lorsque } l_{t+1} \neq j_{t+1}, t = 1, 2, 3$$

$$q_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_t = l_t | Y_t = j_t, G = g)$$

$$\text{lorsque } l_{t+1} = j_{t+1} \text{ et } t = 1, 2, 3$$

$$(A9)$$

$$j_t, l_t \text{ et } m_t \text{ dans le cas de O, C et H}$$

$$q_{m_t j_2}^{m_t j_2} = P(W_2 = m_t | Y_2 = j_2)$$

$$(B6)$$

$$q_{m_t j_1 j_2}^{m_t j_1 j_2} = P(W_1 = m_t | Y_1 = j_1, Y_2 = j_2)$$

$$(B5)$$

$$q_{j_6}^{j_6} = P(Y_6 = l_6 | Y_6 = j_6)$$

$$(B4)$$

$$t = 2, 3, 4, 5$$

$$(B3)$$

$$q_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_t = l_t | Y_t = j_t, Y_{t+1} = j_{t+1})$$

$$(B2)$$

$$t = 2, 3, 4, 5$$

$$\pi_{j_1 j_2 j_3}^{j_1 j_2 j_3} = P(Y_t = j_t | Y_{t-1} = j_{t-1})$$

$$(B1)$$

$$\pi_{j_1}^1 = P(Y_1 = j_1)$$

1) Décomposition de modèle de base

Spécification finale de modèle pour les données de la
SIP, en ce qui a trait à la décomposition de modèle
de base et au patron des restrictions imposées aux
paramètres

ANNEXE B

$$(A15) \quad q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3}$$

$$(A14) \quad q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3}$$

$$(A13) \quad q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3}$$

$$(A12) \quad q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3} = q_{j_2 j_3}^{j_2 j_3}$$

$$(A11) \quad \text{lorsque } m_{t+1} = j_{t+1} \text{ et } t = 1, 2, 3$$

$$q_{m_t j_1 j_2 j_3}^{m_t j_1 j_2 j_3} = P(W_t = m_t | Y_t = j_t, G = g)$$

$$(A10) \quad \text{lorsque } m_{t+1} \neq j_{t+1} \text{ et } t = 1, 2, 3$$

$$= P(W_t = m_{t+1} | Y_t = j_t, W_{t+1} = m_{t+1}, Y_{t+1} = j_{t+1}, G = g) = 1$$

$$q_{m_t j_1 j_2 j_3}^{m_t j_1 j_2 j_3}$$

Le tableau B1, dans l'annexe B, montre le patron des restrictions imposées aux probabilités relatives aux réponses (a) à e)); le tableau indique quels sont les paramètres qui sont établis sur un pied d'égalité et lesquels sont établis à 0, afin d'introduire dans le modèle les contraintes précitées, définies par les équations (B1) à (B6). Le modèle final a été choisi après avoir comparé plusieurs modèles, comme le montre le tableau 5.

Choix du modèle (ME = mobilité estimée)					
Tableau 5					
MODÈLE	L^2	dl	ΔL^2	valeur de	
				p , test	ME
A	2509,5759	2124	5,424	0	5,798
A1	3450,1716	2154	940,5957		
A2	3849,9470	2178	399,7754	0	5,798
B	816,1620	2076	5,888	0,01	5,818
B1	855,2282	2094	39,0662		
B2	864,9657	2106	9,7375		
B3	879,5996	2121	14,6339		
				0,10	6,252

Nous avons commencé l'analyse en estimant un modèle fondé sur l'hypothèse d'erreurs de classification indépendantes (ECI) (modèle A dans le tableau), qui, comme prévu, présente un mauvais ajustement. Les modèles qui suivent sont fondés sur l'étude de Magnac et Visser (1995). Ces auteurs considèrent des transitions mensuelles au cours d'une période plus longue que la nôtre (de janvier 1989 à mars 1992), mais dans le même échantillon de personnes. Ils présupposent que la situation par rapport au marché du travail qui existe dans le mois de l'interview est correctement déclarée, tandis que la probabilité de commettre des erreurs augmente avec l'écart qui sépare le mois de référence et l'époque de l'interview, d'après une fonction déterministe du temps. Les probabilités de réponse sont présupposées constantes tout au long des vagues de l'enquête, et l'on présuppose que les transitions réelles suivent une chaîne de Markov stationnaire de premier ordre. Notre modèle A1 est une version du modèle de Visser et Magnac assujettie à moins de contraintes (aucune hypothèse de stationnarité n'est émise) et qui est appliquée à des transitions trimestrielles relatives à la

Le modèle B introduit la corrélation entre les erreurs de classification en laissant dépendre chaque indicateur de la transition réelle qui a eu lieu entre les époques t et $t + 1$; en outre, il inclut la contrainte (a). La qualité de l'ajustement augmente de manière frappante (voir L^2). Tous les modèles subséquents sont subordonnés au modèle B et des restrictions supplémentaires peuvent être évaluées par un test conditionnel. Le modèle B1 introduit des contraintes sous b), et le modèle B2 les contraintes supplémentaires sous c); le modèle B3 est notre modèle final.

Le tableau 6 présente les taux de transition estimés à l'aide de notre modèle le mieux ajusté. Le marché du travail français est corrigé dans le sens d'une plus grande mobilité. La mobilité moyenne estimée est de 6,252 %. En outre, les probabilités de réponses estimées montrent un patron cohérent par rapport à l'idée selon laquelle la probabilité de commettre des erreurs augmente avec la longueur de la période pour laquelle le répondant doit faire appel à sa mémoire.

REMERCIEMENTS

Les recherches pour le présent article ont été financées par les subventions CNR n° 94.02242.CT10 et MURST n° 02.09.02.110 et n. 02.09.02.124. Nous remercions Michael Visser et Thierry Magnac de nous avoir fourni des données anonymées tirées de l'enquête française sur la population active. Une version préliminaire du présent article a été présentée lors de la Réunion satellite de l'AISE/AISO sur les études longitudinales, qui a eu lieu à Jérusalem, du 27 au 31 août 1997, où nous avons bénéficié de commentaires et de discussions. Nous remercions particulièrement deux examinateurs pour leurs critiques fouillées et leurs suggestions.

Tableau 6
Taux de transition trimestriels observés (×100), EPPA, décembre 1990 à mars 1992
(ME = mobilité estimée)

OO	OC	OH	CO	CC	CH	HO	HC	HH	ME
D90-M91	94,85	4,48	0,67	12,70	66,28	21,02	1,09	1,55	97,36
M91-J91	95,65	1,37	2,98	28,43	62,35	9,22	3,61	1,48	94,91
J91-S91	93,71	4,25	2,04	14,88	82,50	2,62	4,11	3,49	92,40
S91-D91	98,32	1,67	0,01	15,42	83,75	0,83	3,80	0,47	95,73
D91-M92	93,23	5,02	1,75	9,99	88,65	1,36	2,07	1,28	96,65
								5,56	4,24
								7,70	7,49
								6,27	6,27

Tableau 4
Taux de transition trimestriels observés (×100), EFPA, décembre 1990 à mars 1992
(MO = mobilité observée)

	OO	OC	OH	CO	CC	CH	HO	HC	HH	MO
D90-M91	IV	4,25	0,98	24,53	72,40	3,07	0,98	0,66	98,36	5,08
	EV	4,86	3,64	31,60	56,84	11,56	4,40	2,10	93,50	10,16
M91-J91	IV	3,02	0,95	23,21	74,32	2,47	1,28	0,68	98,04	4,54
	EV	4,63	3,89	35,01	54,20	10,79	4,84	2,14	93,02	12,04
J91-S91	IV	3,94	1,77	20,93	78,29	0,78	4,71	2,95	92,34	7,85
	IV	4,48	1,79	23,63	74,89	1,48	3,22	1,65	95,13	7,23
S91-D91	IV	4,48	1,79	23,63	74,89	1,48	3,22	1,65	95,13	7,23
D91-M92	IV	4,80	1,30	21,67	76,74	1,59	1,70	0,59	97,71	5,91

La seule façon de rendre compte d'erreurs de classification présentes dans les données de l'EFPA consiste à laisser dépendre les situations observées de transitions latentes. Par ailleurs, cela semble être une présupposition sensée dans le cas d'enquêtes rétrospectives. En effet, les flux entre deux situations différentes peuvent facilement faire l'objet d'un classement erroné, parce que dans certaines circonstances, il peut vraiment être difficile de situer correctement les événements. Par exemple, des employés qui perdent leur emploi ou qui prennent leur retraite (flux OC et CH) vont généralement prendre les jours de congé auxquels ils ont droit, et peuvent ne pas savoir exactement à quel moment ils ont perdu ou quitté leur emploi. Il peut être également difficile de se rappeler à quel moment on s'est joint à la population active, surtout si l'arrivée sur le marché du travail a eu lieu après avoir quitté l'école (flux HC et HO)(van de Pol et Langeheine 1997).

Le modèle LISREL modifié, formulé mathématiquement dans l'annexe B, est fondé sur les hypothèses de fond énoncées ci-après.

Au niveau latent, les transitions suivent une chaîne de Markov non stationnaire de premier ordre (équations (B1) et (B2)). En effet, l'importance de la saisonnalité dans le cas des transitions observées incite à éviter d'imposer une stationnarité de quelque ordre que ce soit à la chaîne de Markov latente.

Les probabilités relatives aux réponses dans le cas de données recueillies au cours des deux vagues dépendent de la transition latente qui a lieu entre t et $t + 1$ (les équations (B3) et (B4) renvoient à des données recueillies en mars 1992, et les équations (B5) et (B6) à des données recueillies en mars 1991).

Afin de décrire en détail le mécanisme qui est à l'origine des erreurs et de spécifier un modèle plus parcimonieux, nous avons imposé les contraintes suivantes aux probabilités relatives aux réponses:

- a) les probabilités de réponse qui ont trait au même mois faisant partie d'années subséquentes (décembre et mars) sont considérées comme étant égales;
- b) les probabilités de réponse à l'époque t , étant donné que la situation réelle n'a pas changé entre l'époque t et l'époque $t + 1$, sont considérées comme étant constantes au fil du temps;

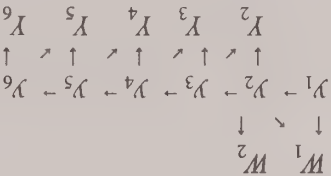


Figure 5. Schéma de parcours d'un modèle LISREL modifié, avec six points de mesure et deux indicateurs pour une seule variable latente

- c) les probabilités de réponse sont considérées comme étant égales pour juin et septembre 1991;
- d) en général, les répondants qui changent de situation entre le mois t et $t + 1$ (transitions OC, OH, CO, CH, et HC) déclarent, à l'époque t , soit la situation réelle à l'époque t , soit la situation réelle à l'époque $t + 1$; autrement dit, les répondants ne déclarent pas une situation qu'ils n'ont pas quitté ou à laquelle ils n'ont pas accédé;
- e) cependant, si la transition latente a lieu entre les situations H et O, nous admettons les trois réponses à l'époque t , c'est-à-dire que nous considérons que les gens qui trouvent un emploi peuvent confondre leur situation précédente (à l'époque t) et hésiter entre C et H.

La contrainte c) est imposée principalement pour obtenir un modèle parcimonieux. Cette contrainte intègre l'idée selon laquelle les probabilités de réponse relatives à des mois qui se situent plus ou moins dans la partie centrale de la période de référence ne varient pas beaucoup. Les contraintes b) et d) reflètent le fait que les probabilités de réponse dépendent de transitions latentes. Nous nous attendons à ce que ces probabilités ne varient pas beaucoup au fil du temps lorsqu'il n'y a pas de changement latent (contrainte b)), tandis que nous prévoyons que la probabilité de mal situer des changements, surtout dans le cas de situations ambiguës, augmente avec la longueur de la période pour laquelle le répondant doit faire appel à sa mémoire. Les contraintes d) visent à tenir compte de l'effet de télescopage.

La figure 5 montre le schéma de parcours du modèle estimé.

en grande partie être élargies aux données de l'enquête française.

Tableau 3

Transitions observées dans le cadre de l'EFPa (×100),

de février à avril 1991

	OO	OC	OH	CO	CC	CH	HO	HC	HH
F-M IV	98,19	1,67	0,14	9,11	90,65	0,24	0,28	0,11	99,61
EV	93,17	3,58	3,25	25,18	65,23	9,59	3,75	1,96	94,29
M-A IV	98,60	1,04	0,36	8,89	90,37	0,74	0,24	0,29	99,47
EV	93,24	3,33	3,43	25,90	63,79	10,31	3,79	2,07	94,14

En général, dans le cas d'une enquête rétrospective comportant une période aussi longue pour laquelle les répondants doivent faire appel à leur mémoire, nous attendons à ce que le manque de mémoire soit la cause principale des erreurs de classification. Nous nous attendons également à ce que la probabilité de donner une réponse incorrecte augmente à mesure que s'accroît l'écart entre le mois de référence et le mois où a lieu l'interview. Ce fait peut être considéré comme la principale cause de corrélation entre erreurs de classification, conjointement avec le télescopage et les effets de conditionnement, qui pourraient également toucher les données de l'enquête française (voir Magnac et Visser 1995).

On peut raisonnablement penser que l'effet global des erreurs de classification entraîne une sous-estimation de la mobilité en ce qui a trait au marché du travail français. Sur la base de ces considérations, nous avons spécifié un modèle dans le but de supprimer les erreurs de mesure dans les flux bruts observés à intervalles trimestriels (tableau 4). La dernière colonne du tableau 4 indique le pourcentage observé de personnes qui ont changé de situation entre les deux mois considérés (MO = mobilité observée). En moyenne, au cours des cinq transitions IV, on observe 6,122 % de mobilité entre deux mois consécutifs.

Comme dans les études de cas précédentes, nous désignons par y_i ($i = 1, 2, 3, 4, 5, 6$) les situations réelles par rapport au marché du travail, et au moyen de lettres majuscules, Y_i ($i = 2, 3, 4, 5, 6$), leurs indicateurs, qui représentent les situations d'activité observées en mars 1992 (en se référant à mars, juin, septembre et décembre 1991, ainsi qu'à mars 1992); W_i ($i = 1, 2$) représente les situations d'activité observées en mars 1991 (en se référant à décembre et mars 1991). Comme auparavant, y_i , Y_i et W_i renvoient à l'une des trois catégories habituelles (O, C et H).

Le modèle est spécifié en décomposant la proportion indiquée dans la cellule générique du tableau de contingence à sept caractères, comme dans l'annexe B, équations (B1) à (B6).

Etant donné que nous observons deux indicateurs pour un mois seulement, un modèle dans lequel on présuppose des effets directs entre les indicateurs serait sous-identifié. C'est pourquoi nous ne sommes pas en mesure de modéliser explicitement la dépendance entre les situations observées.

Aux fins de notre analyse, nous avons regroupé ces huit catégories sous les trois catégories habituelles (O, C, H). Nous considérons comme des personnes occupées les répondants qui se classent dans les trois premières catégories, comme des personnes en chômage les répondants qui se classent dans la quatrième catégorie, et comme personnes hors du marché du travail les répondants qui se classent dans les catégories restantes.

Nous analysons les informations qui ont été recueillies au cours des deux vagues consécutives de mars 1991 et de mars 1992, après d'un sous-échantillon de personnes, c'est-à-dire les personnes qui ont répondu à trois interviews consécutives (janvier 1990, mars 1991 et mars 1992) et qui étaient âgées entre 18 et 20 ans en 1992, pour un total de 5 427 personnes. Les périodes de référence des deux vagues que nous considérons se chevauchent en mars 1991. Nous avons donc deux renseignements sur la situation d'activité pour le mois en question: un qui a été recueilli en mars 1991 et l'autre qui a été obtenu au moyen d'une question rétrospective, douze mois plus tard.

Le patron de transitions mensuelles observées que présente notre échantillon de l'EFPa révèle quelques faits intéressants, qui sont attribuables en grande partie au jeune âge des répondants.

Les transitions montrent un degré modéré de variation saisonnière liée au calendrier scolaire. De juin à juillet, par exemple, nous observons une proportion plus grande que la moyenne de gens qui arrivent sur le marché du travail comme personnes occupées; par contre, d'août à septembre, une proportion de gens plus grande que la moyenne quitte la situation d'emploi (vraisemblablement pour étudier).

La distribution marginale des trois situations entre mars 1990 et mars 1992 montre que les personnes faisant partie de notre échantillon arrivent progressivement sur le marché du travail: en mars 1990, on constate que 44 % d'entre elles occupent un emploi ou sont au chômage, tandis qu'en mars 1992, cette proportion a augmenté à 54 %.

Le renseignement double pour le mois de mars 1991 fournit quelques preuves concernant les erreurs de réponse contenues dans les données: 8 % des répondants déclarent une situation différente dans l'une des interviews par rapport à l'interview précédente. Dans le cas de la période de février à avril 1991, on observe deux types de flux: un flux intra-vague (IV), c'est-à-dire que les informations sur la situation par rapport au marché du travail ont été recueillies au cours d'une même interview, et un flux entre vagues (EV), c'est-à-dire que les informations ont été recueillies au cours de deux interviews distinctes (tableau 3).

Comme prévu, les transitions IV dénotent un marché du travail plus stable que les transitions EV. Ce fait peut être considéré comme une indication de la présence d'erreurs de classification corrélées dans les données. Nous avons abordé de manière approfondie les patrons et les causes de la corrélation entre erreurs dans les deux sections précédentes, et les considérations qu'elles contiennent peuvent

1986 (tableau 2). La comparaison entre les flux observés et les flux estimés met en évidence le fait que le modèle réduit l'effet de lisière: les transitions IV sont corrigées dans le sens d'un marché du travail plus dynamique, tandis que les transitions EV sont corrigées dans le sens opposé. Il est utile de noter que les effets de la correction par modélisation sont plus évidents dans le cas des flux qui partent de la situation «sans emploi», qui sont caractérisés par une plus grande mobilité.

La qualité de l'ajustement du modèle a été évaluée au moyen de critères multiples, comme le CIB et le test conditionnel pour modèles hiérarchiques, ainsi qu'en faisant appel à l'interprétabilité et à la cohérence par rapport à des connaissances de fond sur la dynamique du marché du travail américain des années 80.

4.3 Les données de l'enquête française sur la population active

La deuxième étude de cas porte sur les flux observés sur les marchés du travail dans le cadre de l'enquête française sur la population active (EFPA), qui est effectuée annuellement par l'INSEE.

La population de référence de l'EFPA est composée de tous les membres des ménages français qui sont âgés de quinze ans et plus dans l'année au cours de laquelle doit avoir lieu l'interview. Le plan d'échantillonnage de cette enquête prévoit tous les ans le renouvellement d'un tiers de l'échantillon. Les informations sur l'activité sont recueillies au moyen de questions rétrospectives qui portent sur la période de référence de treize mois qui a précédé l'interview. À chaque répondant, on demande d'indiquer, sur une base mensuelle, quelle était sa situation par rapport au marché du travail durant la période de référence; pour ce faire, le répondant dispose d'une grille où il ou elle peut indiquer sa situation au cours de chaque mois. Cette grille comporte huit catégories: indépendant, occupé pour une durée déterminée, occupé de façon permanente, sans emploi, en formation, étudiant(e), service militaire et autre (retraité(e), ménagère, etc).

Tableau 2
Transitions mensuelles observées et estimées dans le cadre de la SIPP
(×100), janvier à avril 1986

	OC	OC	OH	CO	CC	CH	HO	HC	HH
J-F	IV	98,11	1,17	0,72	14,53	80,16	5,31	0,90	1,57
	EV	94,08	2,17	3,75	23,58	44,30	32,12	5,62	3,45
	Estimé	97,25	1,47	1,28	16,08	77,16	6,76	1,59	1,32
F-M	IV	98,66	0,92	0,42	16,06	78,67	5,27	0,64	1,65
	EV	94,88	1,91	3,21	21,90	48,54	29,56	4,99	4,11
	Estimé	97,83	1,20	0,97	19,40	74,01	6,59	1,21	1,50
M-A	IV	98,71	0,64	0,65	20,76	71,74	7,50	1,47	1,05
	EV	95,59	1,52	2,89	30,48	34,92	34,60	6,34	3,78
	Estimé	98,11	0,95	0,94	26,42	65,75	7,83	2,17	0,71

Ces contraintes sont formulées en détail dans l'annexe A. Essentiellement, elles intègrent une connaissance *a priori* du comportement des répondants et nous permettent de spécifier un modèle parcimonieux. En particulier, les équations (A8) à (A14) correspondent aux énoncés suivants:

- a) concernant les erreurs de classification IV, on pré-suppose, en suivant Hubble et Judkins (1989),
- a1) qu'un répondant qui déclare une situation d'activité erronée va continuer à donner la même réponse pour le mois qui suit immédiatement, en remontant dans la vague (A8);
- a2) cependant, si la situation à l'époque $t + 1$ est déclarée correctement, la probabilité de réponse pour le mois qui précède immédiatement dépend uniquement de la situation réelle actuelle (A9);
- a3) le même mécanisme à l'origine d'erreurs agit dans le cas des deux indicateurs. Dans le cas de W_t , nous énonçons qu'une réponse correcte est donnée lorsque la situation réelle est O et qu'un emploi est déclaré, et lorsque la situation réelle est C ou H, et qu'aucun emploi n'est déclaré, (A10) et (A11).
- b) Les probabilités de réponse établies sont les mêmes dans le cas de tous les groupes de probabilités, (A12) à (A15). Par exemple, les égalités dans (A12) signifient que les probabilités de réponse, dans le cas des personnes faisant partie du groupe de renouvellement 1, pour le mois d'avril, sont égales aux probabilités de réponse relatives aux personnes faisant partie du groupe 4, pour le mois de mars, aux probabilités relatives au groupe 3, pour le mois de février, et aux probabilités relatives aux personnes du groupe 2, pour le mois de janvier. (Les probabilités sont établies de manière à ce qu'elles soient égales car elles renvoient à la réponse donnée pour le dernier mois de la vague.)
- Le modèle a été estimé pour corriger les flux bruts observés relatifs au trimestre compris entre janvier et avril

l'histoire prédomine sur les autres sources d'erreurs qui peuvent biaiser les estimations de flux bruts. En récapitulant, le recours à une méthode fondée sur la modélisation pour obtenir des flux bruts non biaisés à partir de données de la SIPP est justifié par les deux arguments suivants :

- a) la présence manifeste d'erreurs de classification corrélées;
- b) le recours à des informations a priori sur le mécanisme à l'origine des données tirées de deux sources; b1) des preuves précises tirées des flux bruts observés de la SIPP, comme l'effet de lisière et l'accroissement de la stabilité lorsqu'on recule à l'intérieur de la vague (mentionnées précédemment);
- b2) des explications générales contenues dans la documentation sur les enquêtes portant sur le comportement des répondants.

Afin de supprimer, dans les données de la SIPP, les erreurs de classification contenues dans les flux bruts relatifs à la population active, nous avons élaboré un modèle d'après les hypothèses et les informations suivantes :

- a) le processus de transition réelle suit une chaîne de Markov de premier ordre;
- b) des données sur les transition IV entachées d'erreurs de classification corrélées, d'après un patron qui est spécifique ci-après;
- c) dans le cas des transitions EV, l'hypothèse d'erreurs de classification indépendantes est confirmée;
- d) les groupes de renouvellement sont des échantillons équivalents aux fins d'une modélisation également, c'est-à-dire que les répondants se comportent de la même manière dans les quatre groupes de renouvellement;
- e) les données de la SIPP donnent deux indications sur la situation mensuelle par rapport au marché du travail de chaque répondant: les informations détaillées recueillies dans la section du questionnaire intitulée Labour Force and Reciprocity, dont il a été question précédemment, ainsi que les renseignements supplémentaires recueillis dans la section intitulée Earnings and Employments, dans laquelle on demande aux répondants, sur une base hebdomadaire, s'ils occupaient ou non un emploi au cours de la période de référence.

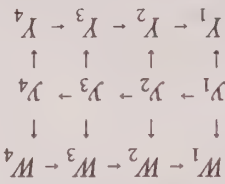


Figure 4. Schéma de parcours d'un modèle LISREL modifié pour quatre points de mesure et deux indicateurs pour chaque variable latente (pour connaître la signification des symboles, voir le texte principal)

La figure 4 représente le schéma de parcours d'une version simplifiée du modèle (c'est-à-dire une version qui n'a pas pour but de représenter en détail le patron d'erreurs de classification corrélées, ni de prendre en considération le fait que nous sommes en présence de quatre groupes de renouvellement) pour quatre points temporels, c'est-à-dire quatre mois consécutifs. Dans ce cas, $y_t (t = 1, 2, 3, 4)$ représente des variables latentes, Y_t et W_t représentent des fichiers représentant des effets directs entre paires de variables. L'indicateur Y_t renvoie à la situation d'activité déclarée, qui est décrite au moyen des trois catégories habituelles (O, C et H), tandis que W_t renvoie à la variable binaire «emploi/aucun emploi». Compte tenu du fait que les informations sont recueillies dans deux sections distinctes du questionnaire, et par des méthodes d'interview différentes, on peut présupposer qu' Y_t et W_t sont indépendantes, étant donné y_t . D'autre part, les effets directs entre les indicateurs sont à l'origine d'erreurs de classification corrélées au fil du temps; la réponse donnée pour l'époque $t + 1$ a une incidence sur celle qui est donnée pour l'époque t . Il est à noter également qu'une variable supplémentaire G , qui comporte quatre catégories, devrait être ajoutée au schéma, afin de rendre compte du nombre de personnes faisant partie des groupes de renouvellement. Tous les indicateurs dépendent de G , étant donné que les unités faisant partie de divers groupes sont interviewés au cours de mois civils différents. L'équation de base du modèle décompose la proportion contenue dans la cellule générique du tableau de contingence à neuf caractères pour l'obtention du produit des probabilités conditionnelles énoncées dans l'annexe A, équations (A1) à (A7). Une version préliminaire du modèle a été proposée par Bassi, Croon, Hageenaars et Vermunt (1994).

L'équation (A1) définit la probabilité d'appartenir à l'un des quatre groupes de renouvellement. Les équations (A2) et (A3) définissent respectivement la situation initiale et les probabilités de transition de la chaîne de Markov de premier ordre latente. Les équations (A4) et (A5) définissent les probabilités de réponse concernant l'indicateur Y_t ; enfin, les équations (A6) et (A7) définissent les probabilités analogues concernant l'indicateur dichotomique W_t . Les probabilités de réponse sont définies de manière à ce que la réponse qui est donnée pour un mois donné dépende à la fois de la situation réelle actuelle (y_t) et des situations «passées» réelles et «passées» déclarées (y_{t+1} et Y_{t+1}). Le mot «passée» renvoie à la façon de penser des répondants lorsqu'ils répondent à des questions rétrospectives: ils commencent à se souvenir à partir du moment le plus proche de l'interview et remontent jusqu'à la fin de la période de référence.

Un ensemble complexe de contraintes a été imposé aux probabilités de réponse de (A4), (A5), (A6) et (A7), afin de tenir compte, premièrement, de l'effet de conditionnement, et, deuxièmement, du fait que les quatre groupes de renouvellement sont des échantillons équivalents en ce qui trait au mécanisme qui est à l'origine des erreurs.

Mois	Groupe de		Mois de référence	
d'interview	renouv.	Vague		
Février	2	1	Oct	Nov
Mars	3	1	Nov	
Avril	4	1	Déc	Jan
Mai	1	1	Fév	
Jun	2	2	Mars	Avr
Jullet	3	2	Avr	Mars
Août	4	2	Mai	Avr
Septembre	1	2	Juin	Mai

Figure 3. Plan de renouvellement pour le panel de 1986 de la SIPP (deux premières vagues)

Les flux bruts observés entre deux mois civils génériques sont alors obtenus de la manière énoncée ci-après.

- a) Dans le cas des personnes appartenant aux trois premiers groupes de renouvellement, d'après des données rétrospectives recueillies lors d'une même interview. Ces flux observés sont appelés transitions «intra-vague» (IV).
- b) Dans le cas des personnes appartenant au quatrième groupe de renouvellement, en réunissant les informations recueillies lors de deux interviews différentes, effectuées à quatre mois d'intervalle. Ces flux observés sont appelés transitions «entre vagues» (EV).

Dans l'estimation des variations mensuelles, un problème particulier que l'on rencontre dans le cas des données de la SIPP est l'effet dit «de lisière» (Young 1989): on observe davantage de variations lorsque les données relatives à deux mois qui se suivent sont recueillies dans deux vagues différentes (la transition chevauche la lisière entre les deux vagues) que dans le cas où les données proviennent d'une même interview. L'effet de lisière est présent dans l'ensemble de l'enquête; des preuves de cet effet en ce qui a trait à plusieurs variables importantes sont décrites par Martini (1988), Marquis et Moore (1989), ainsi que par Kalton et Miller (1991).

Le tableau 1 illustre ce phénomène dans le cas de notre échantillon du panel de 1986 de la SIPP. La ligne 4-1 contient les transitions IV moyennes; les lignes 1-2, 2-3 et 3-4 contiennent les transitions EV moyennes relatives à la position des deux mois de référence pertinents dans chaque vague (par exemple, la ligne 1-2 contient les transitions qui ont eu lieu dans chaque vague entre les deux premiers mois de référence). Le tableau 1 montre clairement que les transitions IV observées décrivent un marché du travail beaucoup plus stable que les transitions EV. De plus, la stabilité IV augmente à mesure qu'on recule dans la vague (de 3-4 à 1-2).

Une cause plausible de l'origine de l'effet de lisière et du patron systématique de transitions observées dans l'en- semble d'une vague est la différence entre l'incidence des erreurs de mesure contenues dans les données recueillies selon la stratégie IV et l'incidence des erreurs de mesure

touchant les données obtenues selon la stratégie EV. Plus précisément, il est probable que les erreurs de classification de flux observés (IV ou EV). La stabilité plus grande que reflètent les transitions IV pourrait être due à une corrélation élevée entre les erreurs de classification. En effet, s'il n'y avait pas de corrélation entre les erreurs, en particulier dans le cas des transitions IV, on ne s'attendrait pas à observer des preuves de l'existence d'un effet de lisière. Un grand nombre de causes plausibles des erreurs corrélées sont évoquées dans les textes spécialisés portant sur la psychologie cognitive et sur les techniques d'enquête, en particulier sur les effets des erreurs de mémoire (voir à cet égard Bernard, Killworth, Kronenfeld et Sailer 1984, ainsi que O'Muircheartaigh 1996); parmi ces causes, on mentionne un effet de «conditionnement»: les répondants ont tendance à donner la même réponse lorsqu'on recule à l'intérieur de la vague, et dans des cas extrêmes, ils répètent machinalement la même réponse pour les quatre mois.

Tableau 1

Transitions mensuelles (×100) dans le cas du panel de 1986 de la SIPP, janvier 1986 à janvier 1987

Type	OO	OC	OH	CO	CC	CH	HO	HC	HH
1-2	IV	98,27	1,04	0,69	15,46	79,63	4,91	1,15	1,42 97,43
2-3	IV	97,91	1,13	0,96	17,34	75,96	6,70	1,38	1,71 96,91
3-4	IV	97,85	1,20	0,95	19,23	73,25	7,52	1,28	1,69 97,03
4-1	EV	94,03	2,10	3,87	26,81	42,20	30,99	5,65	3,77 90,58

Une abondante documentation empirique montre que ce genre d'effet de conditionnement est la principale source d'erreurs de classification dans le cas des données de la SIPP. D'autres sources d'erreurs potentielles, typiques des enquêtes par panel, n'ont pas d'incidence importante sur les données de la SIPP. Les études de vérification de dossiers administratifs fournissent peu de preuves, s'il en existe, d'un effet lié au temps de présence dans l'échantillon (Chakrabarty et Williams 1989; McCormick, Butler et Singh 1992). Comme considération d'ordre général, nous pouvons dire que dans les données de la SIPP, l'effet de

4. DEUX ETUDES DE CAS

4.1 Le cadre général

Dans la présente section, nous présentons deux applications du modèle LISREL modifié qui servent à corriger des flux bruts observés relatifs au marché du travail. Les données utilisées proviennent des deux enquêtes mentionnées ci-après, qui sont fondées sur des plans d'échantillonnage en partie différents:

- 1) la U.S. Survey of Income and Program Participation (SIPP), enquête-ménage par panels multiples qui recueille des informations rétrospectives sur les antécédents d'activité entre vagues;
- 2) l'enquête française sur la population active (EFPA), enquête rétrospective annuelle comportant des périodes de référence qui se chevauchent sur un mois.

Pour chacune des études de cas, nous spécifions un modèle sur la base d'informations a priori qui ont trait tant au processus de transition qu'au mécanisme à l'origine des erreurs. Dans la spécification des modèles, les informations a priori sont essentielles à l'obtention de modèles parcimonieux et plausibles.

Tous les modèles sont écrits sous forme d'un modèle LISREL modifié et estimés à l'aide d'un algorithme à mémoire externe. Nous avons utilisé le programme /BM (Vermunt 1993) et vérifié, dans le cas de tous les modèles, les maximums locaux.

Les deux derniers modèles s'avèrent plutôt complexes, étant donné qu'ils intègrent la corrélation existant entre les erreurs de classification et des hypothèses spécifiques sur le comportement des répondants. Ce fait, ainsi que le caractère clairsemé et non équilibré du tableau de contingence observé, typique des transitions relatives à la population active, exige des critères d'évaluation de la qualité d'ajustement autres que L^2 et X^2 . Dans la première étude de cas, des modèles de remplacement ont été évalués au moyen du CIB et sur la base de connaissances de fond relatives au marché du travail américain. Dans le deuxième cas, des modèles de remplacement ont été comparés au moyen du test conditionnel.

4.2 Les données de la SIPP

La SIPP est une enquête-ménage par panels multiples effectuée par le U.S. Bureau of the Census dans le but de recueillir des informations sur des sujets comme l'emploi, le revenu, la participation à des programmes sociaux, etc. La population de référence est composée des personnes non institutionnalisées âgées de plus de quatorze ans. Cette enquête permanente a démarré en 1984. De façon régulière, un nouvel échantillon de ménages, appelé «panel», a été choisi tous les ans et suivi durant une période

de deux ans et demi (pour une description détaillée de la SIPP, voir U.S. Department of Commerce 1991, ainsi que Citro et Kalton 1993). Chaque panel est divisé de façon aléatoire en quatre «groupes de renouvellement» et interviewé huit fois à intervalles de quatre mois. Pour des raisons pratiques, chaque groupe de renouvellement est interviewé une fois par mois au cours d'une période de quatre mois consécutifs, et au moyen de questions rétrospectives, on recueille des informations sur la période de quatre mois qui s'écoule avant la tenue des interviews subséquentes. Chaque série d'interviews effectuées auprès de l'ensemble de l'échantillon est désignée par le terme «vague».

Nous allons nous référer au panel de 1986, qui a commencé au mois de février 1986 et qui a pris fin au mois d'août 1988. Nous allons considérer la période intermédiaire comprise entre janvier 1986 et janvier 1987, au sujet de laquelle nous disposons d'informations provenant de quatre groupes de renouvellement. La figure 3 montre le plan d'échantillonnage de l'enquête en ce qui a trait à notre échantillon.

Les informations sur l'activité sont recueillies principalement dans la section du questionnaire intitulée «Labour Force and Reciprocity» (pour un renseignement supplémentaire, qui est recueilli dans une autre section du questionnaire, voir ci-après), où l'on demande à chaque répondant de déclarer sur une base hebdomadaire ses antécédents d'activité pour la période de quatre mois (dix-huit semaines) précédente, en passant en revue une série de questions filtrées. On demande tout d'abord au répondant s'il ou elle occupait un emploi ou avait une entreprise à un moment quelconque durant la période de référence. Dans le cas d'une réponse négative, on demande au répondant d'indiquer s'il ou elle a passé du temps à chercher du travail ou s'il ou elle était mis(e) en disponibilité, et dans l'affirmative, au cours de quelles semaines exactement. Par contre, si le répondant répond oui à la première question, (c'est-à-dire s'il ou elle a travaillé durant un certain temps) et mentionne un emploi ou une entreprise comportant une certaine durée au cours de la période de référence, il est invité à passer à la section suivante du questionnaire. Aux répondants qui ne mentionnent pas une situation stable par rapport au marché du travail, on pose une longue série de questions, afin de déterminer la situation d'activité dans laquelle ils se trouvaient au cours de chacune des semaines de référence.

Les informations recueillies sur une base hebdomadaire sont généralement consignées pour obtenir une classification mensuelle fondée sur les trois catégories courantes suivantes: personne occupée (O), en chômage (C) et hors de la population active (H). Dans le cas de personnes qui occupent différentes catégories au cours d'un même mois, la situation d'activité mensuelle est celle identifiée par la catégorie «modale» en ce qui a trait aux semaines du mois en question (Martini 1989).

qu'il existe une source d'association supplémentaire entre les indicateurs qui se situe au-delà de la partie qui est expliquée par le rapport existant entre ces indicateurs et les

variables latentes.

Une fois qu'un modèle raisonnable a été spécifié, il faudrait vérifier l'identification. Le modèle comporte un grand nombre de variables non observables, et l'identification de tous les paramètres n'est pas assurée automatiquement.

Deux stratégies, utilisées éventuellement conjointement,

offrent des possibilités raisonnables de réaliser l'identification :

i) l'imposition de restrictions d'égalité plausibles

à l'ensemble de paramètres; ii) le recours à des indicateurs

multiples de la situation réelle non observée. Le modèle de

Markov à classe latente représentée dans la figure 1, par

exemple, n'est pas identifié sans que l'on n'impose des

restrictions supplémentaires à ses paramètres. Si l'on

présume que la chaîne latente est homogène au fil du temps,

ou si les probabilités de réponse font l'objet de restrictions

pour qu'elles soient égales au fil du temps, on peut montrer

que le modèle est identifié (Lazarsfeld et Henry 1968). La

disponibilité d'indicateurs multiples relatifs à la situation

réelle non observée peut également aider à identifier des

modèles de mesure complexes. Les critères d'identification

utilisés pour certaines spécifications très particulières se

sont avérés efficaces (par exemple, on peut montrer que le

modèle représenté dans la figure 2 est identifié, mais des

règles générales permettant de vérifier l'identification

globale n'ont pas encore été proposées. Il est conseillé de

vérifier au moins l'identification locale, c'est-à-dire l'iden-

tifiabilité des paramètres inconnus dans un voisinage de la

solution du maximum de vraisemblance. Selon Goodman

(1974), une condition suffisante pour permettre l'identi-

cation d'un modèle à classe latente est que la matrice

d'information soit à rang complet. La condition énoncée

par Goodman peut être difficile à vérifier du point de vue

du calcul. En outre, dans le cas de certains ensembles de

données, il peut arriver que la matrice d'information ne soit

pas à rang complet, simplement parce que certaines estima-

tions sont très proches des bornes de l'espace des para-

mètres. Une autre façon, empirique, de vérifier l'identi-

fiabilité consiste à estimer le modèle en utilisant divers

ensembles de valeurs de départ. Si des ensembles de valeurs

de départ différents donnent la même valeur pour la fonc-

tion de vraisemblance logarithmique, mais des estimations

de paramètres différentes, le modèle n'est pas identifié.

Pour ce qui est de l'estimation, les modèles LISREL

peuvent être traités comme des modèles log-linéaires

dirigés à variables latentes (Hagenaars 1997). Un modèle

log-linéaire dirigé donne une séquence de modèles logit

multinomiaux partitionnés (comportant éventuellement

des variables latentes) qui sont estimés par étapes. À

chaque étape, on considère une variable dépendante et on

estime un modèle logit multinomial sur un tableau de

contingence qui a été réduit au-dessus des variables qui

n'ont pas d'incidence directe sur la variable dépendante

dans l'ordre causal. Finalement, les estimations obtenues à chaque étape sont réunies, afin d'obtenir des paramètres estimés pour le modèle complet. La modélisation log-linéaire dirigée donne exactement les mêmes estimations, les mêmes erreurs-types et les mêmes statistiques de contrôle que la méthode standard de Goodman, mais en utilisant des tableaux marginaux plus simples. Si le modèle causal contient une ou plusieurs variables latentes, il faut utiliser une méthode d'estimation appropriée, par exemple, une application de l'algorithme à mémoire externe (Meng et Rubin 1993).

La validité empirique du modèle causal complet peut être

testée en comparant les fréquences prévues estimées avec

les fréquences observées, dans le tableau complet, à l'aide

du rapport de vraisemblance L^2 et de la statistique de

Pearson X^2 . Cependant, en raison de la nature de la

structure des données observées relatives aux transitions qui

ont lieu sur le marché du travail, de nombreuses cellules

indiquent des fréquences observées très faibles. Pour cette

raison, les critères habituels X^2 et L^2 devraient être utilisés

uniquement pour obtenir une indication générale d'ajuste-

ment, étant donné que leur distribution asymptotique χ^2

n'est plus garantie, en raison de la répartition clairsemée et

non équilibrée des fréquences dans le tableau de

contingence.

Diverses stratégies permettent d'élargir et d'améliorer

l'évaluation des modèles; trois d'entre elles méritent d'être

mentionnées dans le présent contexte:

i) Un modèle restreint subordonné à un modèle plus vaste

peut être vérifié à l'aide du test conditionnel, c'est-à-

dire en considérant la différence existant entre les deux

modèles en ce qui a trait à la valeur de L^2 , qui est

distribuée asymptotiquement comme χ^2 dans des

conditions plus faibles (Goodman 1981, et Haberman

1978).

ii) En général, le recours à des critères multiples peut être

une stratégie sensée. Des indices fondés sur le critère

d'information, comme le critère d'information d'Akaike

(CIA) ou le critère d'information de Bayes (CIB),

peuvent être utiles pour comparer des modèles de

remplacement non hiérarchiques. Un autre avantage

du CIA et du CIB est que, dans la procédure de sélec-

tion, ces critères soupèsent la qualité d'ajustement d'un

modèle par rapport à la partition de celui-ci, en

considérant les degrés de liberté du modèle et la taille

de l'échantillon. (CIA = $L^2 - 2 \times \text{degrés de liberté}$,

CIB = $L^2 - \ln(N+1) \times \text{degrés de liberté}$.) Le modèle

que l'on préfère dans ce contexte est celui qui présente

la plus faible valeur du CIA ou du CIB.

iii) Des méthodes de rééchantillonnage de Monte Carlo

peuvent être mises en oeuvre pour simuler la dis-

tribution asymptotique de X^2 et L^2 (Langheine, Pannekoek et van de Pol 1995).

d'après des études à réinterview peuvent être représentées dans le cadre décrit précédemment. Dans ce cas, on a recours à des informations supplémentaires; c'est-à-dire que les paramètres q_t sont estimés à l'extérieur du cadre d'observation et introduits dans (3.5) afin d'obtenir directement $P(y_1, y_2)$.

Le même cadre peut être utilisé pour inclure des hypothèses plus générales au sujet du processus latent comme du processus de mesure, jusqu'à l'inclusion d'erreurs de classification corrélées en série. Comme cas intéressant, nous considérons le modèle proposé par Pfeffermann, Skinner et Humphreys (1998). Sans tenir compte ici de la situation initiale, voici comment ces auteurs expriment les probabilités de réponse conditionnelles:

$$q_{t,t'}' = P(Y_t = t' | y_t = j_t, Y_{t-1} = l_{t-1}) \quad t = 2, \dots, T, \quad (3.6)$$

ils maîtrisent ainsi l'hypothèse d'erreurs de classification indépendantes.

Une formulation similaire, qui vise à introduire, du moins en partie, une dépendance entre la situation observée à l'époque t et la séquence de situations réelles existant aux époques t et $t-1$, a été proposée par van de Pol et Langeheine (1992), qui élargissent le modèle afin de pouvoir inclure également une chaîne de Markov de deuxième ordre pour le processus de transitions réelles. La stratégie de modélisation servant à estimer les flux bruts peut être élargie davantage dans diverses directions, dont celles décrites ci-après.

- a) Le modèle peut être élargi tout naturellement pour exploiter la disponibilité d'indicateurs multiples d'une même situation réelle non observée. Cela implique que les probabilités de réponse, comme celles contenues dans (3.2), sont définies pour une ou plusieurs autres variables observées supplémentaires, qui sont traitées comme des mesures imparfaites de la même situation latente y_t . La figure 2 montre, à titre d'exemple, un modèle de Markov à classe latente dans le cas de deux indicateurs par variable latente et quatre points temporels. Dans ce modèle, chaque paire d'indicateurs qui renvoie à un point temporel donné est présumée être indépendante, sous condition de la variable latente correspondante; c'est-à-dire que la corrélation entre les indicateurs est complètement expliquée par leur rapport avec y_t .
- b) L'hétérogénéité observée au niveau individuel, dans le processus de transition ou dans le processus de mesure, peut être introduite par l'entremise de conditions relatives à un ensemble de covariables X_t . Un exemple de cette approche est contenu dans l'étude de Pfeffermann, et coll. (1998). Ces auteurs utilisent des informations sur des covariables au niveau de l'unité et modélisent leur incidence sur la situation d'activité à l'aide d'un logit multinomial.

- c) L'hétérogénéité non observée peut également être prise en considération, ce qui mène à des modèles à classe latente mixtes (van de Pol et Langeheine 1990). Un exemple simple de ce genre de modélisation est le modèle des personnes qui déménagent et des personnes qui restent, dans lequel on présuppose un comportement différent, au niveau latent, dans le cas de groupes d'unités, lorsque le nombre d'unités faisant partie d'un groupe ne peut être observé directement.

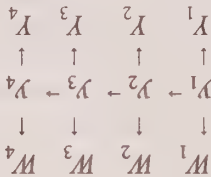


Figure 2. Schéma de parcours d'un modèle de Markov à classe latente dans le cas de quatre points de mesure et deux indicateurs pour chaque variable latente

3.2 Utilisation de modèles à classe latente et de modèles connexes dans l'estimation de flux bruts entachés d'erreurs de mesure

Les modèles à classe latente sont un exemple particulier de la formulation de modèles généraux que nous avons décrite dans la section précédente. Dans ces modèles, la situation d'activité réelle joue le rôle de variable latente, tandis que la situation observée sert d'indicateur de cette variable. Certaines des spécifications mentionnées dans la section précédente comprennent la dépendance entre les erreurs de classification. Une méthode générale et pratique qui permet de traiter cette dépendance, et qui comprend des modèles à classe latente standard avec erreurs de classification corrélées, est le modèle appelé LISREL, modifié qui a été proposé par Hagenaars (1990).

Le modèle LISREL, modifié est un prolongement de l'analyse de parcours de Goodman (1973), qui est un outil servant à décrire de type causal existant entre des variables nominales par l'entremise d'un système d'équations logit. Essentiellement, le prolongement intègre des variables latentes. Ainsi, un modèle LISREL modifié réunit un sous-modèle de mesure, qui spécifie la dépendance des indicateurs par rapport à des variables latentes, et un sous-modèle structurel qui spécifie des rapports ordonnés existant entre des variables latentes et entre d'éventuelles variables extérieures. Comme le suggère son nom, ce modèle peut être considéré également comme l'analogue pour variables discrètes du modèle LISREL bien connu pour variables continues (Joreskog et Sörbom 1988).

Les modèles LISREL modifiés permettent d'introduire des erreurs de classification corrélées en série en insérant des effets directs implique que l'association entre les variables observées n'est pas entièrement expliquée par les effets des variables latentes sur leurs indicateurs, mais

de données rétrospectives, ainsi que d'indicateurs multiples d'une même variable latente, est utilisée pour obtenir des modèles parcimonieux.

3. ESTIMATION DE FLUX BRUTS ENTACHÉS D'ERREURS DE CLASSIFICATION

3.1 Un cadre général

La spécification de modèles statistiques qui est effectuée en vue de corriger des flux bruts de population active pour tenir compte d'erreurs de classification devrait permettre de prendre en considération la nature des données dont on dispose (comme nous l'avons indiqué dans la section précédente) et des hypothèses solides au sujet des processus qui sont à l'origine, premièrement, des transitions entre les situations d'activité (par ex., des chaînes de Markov) et, deuxièmement, des erreurs de mesure (par ex., des erreurs non corrélées et des erreurs corrélées).

Dans le cas le plus simple, nous considérons des données de panel où pour chaque période $t = 1, \dots, T$, on observe une variable discrète Y_t relative à une unité générique faisant partie d'un échantillon aléatoire de taille n . Dans nos études de cas, les unités sont des personnes et les périodes sont des mois ou des trimestres. Y_t représente une valeur ou une situation choisie parmi un nombre r de valeurs ou de situations distinctes. Y_t est une mesure impartiale d' y_t , qui représente la situation réelle d'une unité générique à l'époque t . En général, il n'est pas nécessaire de présupposer qu' y_t varie au cours du même ensemble de situations 1, 2, ..., r , mais afin de simplifier les choses, et sans diminuer la portée générale, nous allons considérer ici le même ensemble de situations que dans le cas d' Y_t .

Les stratégies qui servent à estimer des flux bruts sont fondées sur une spécification appropriée de la probabilité combinée relative aux processus réel et observé $P(Y_1, \dots, Y_T, y_1, \dots, y_T)$. L'analyse statistique s'appuie donc sur la marginalisation par rapport à des grands observés:

$$P(Y_1, \dots, Y_T) = \sum_{y_1=1}^r \dots \sum_{y_T=1}^r P(Y_1, \dots, Y_T, y_1, \dots, y_T). \quad (3.1)$$

Les modèles sont fondés sur des spécifications parcimonieuses de la fonction de probabilité combinée $P(Y_1, \dots, Y_T, y_1, \dots, y_T)$. Cette parcimonie peut être obtenue essentiellement en décomposant la fonction en un produit de probabilités conditionnelles découlant d'un ensemble approprié d'hypothèses concernant la structure de dépendance existant entre les composantes $Y_1, \dots, Y_T, y_1, \dots, y_T$. Pour nos besoins, un point de départ pratique pour la spécification des modèles est offert par des hypothèses formulées, premièrement, au sujet de la structure du processus qui est à l'origine des transitions réelles entre les situations d'activité, et, deuxièmement, au sujet du processus de mesure (en utilisant, par exemple, des connaissances

solides ou des preuves empiriques tirées de la stratégie de collecte de données qui a été adoptée). Dans le cas d'un modèle dont le but est de faire la distinction entre les transitions réelles et observées relatives au marché du travail, un exemple typique d'application de cette idée est fourni par les modèles de Markov à classe latente (van de Pol et Langeheine 1990). Les probabilités suivantes sont spécifiées dans le cas d'une unité générique:

$$q_{t',t}^i = P(Y_t = i | y_t = j_t) \quad t = 1, \dots, T \quad (3.2)$$

$$\pi_{t',t-1}^{j_t} = P(y_t = j_t | y_{t-1} = j_{t-1}) \quad t = 2, \dots, T \quad (3.3)$$

$$\pi_{j_t}^i = P(Y_t = j_t) \quad (3.4)$$

Les probabilités conditionnelles (3.2) représentent le rapport existant entre la situation réelle et la situation observée, c'est-à-dire la probabilité de déclarer, à l'époque t , la situation i alors que la situation réelle est j_t . Il est clair que cette spécification implique l'hypothèse d'indépendance locale, c'est-à-dire que Y_1, \dots, Y_T sont indépendantes, étant donné y_1, \dots, y_T . Les probabilités conditionnelles (3.3) représentent la dynamique du marché du travail, c'est-à-dire la probabilité qu'une transition de j_{t-1} à j_t ait lieu lorsqu'on passe de l'époque $t-1$ à t . D'après (3.3), le processus de transition réel évolue en suivant une chaîne de Markov de premier ordre. Enfin, les probabilités (3.4) décrivent la situation initiale du processus de Markov. La probabilité marginale dans le cas de la séquence observée (3.1) est donnée alors par:

$$P(Y_1 = i_1, \dots, Y_T = i_T) = \sum_{j_1=1}^r \dots \sum_{j_T=1}^r \pi_{j_1}^{i_1} \prod_{t=2}^T q_{t',t-1}^{j_t} \pi_{j_T}^{i_T} \quad (3.5)$$

Dans le cas de quatre points de mesure, le modèle (3.5) est représenté de façon équivalente par le schéma de parcours de la figure 1, où les flèches indiquent des effets directs entre les variables.

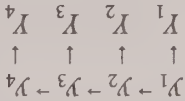


Figure 1. Schéma de parcours d'un modèle de Markov à classe latente dans le cas de quatre points de mesure.

Il est utile de noter que l'hypothèse d'indépendance locale équivaut à l'hypothèse d'erreurs de classification indépendante (BCI). Comme nous l'avons indiqué dans la section précédente, l'hypothèse BCI est sévèrement critiquée et paraît absolument déraisonnable dans le cas d'une collecte de données effectuée au moyen de questions rétrospectives. Comme autre exemple, dans le cas de $T=2$, des stratégies classiques utilisées pour corriger des flux bruts

le cas d'EPA comportant un plan d'échantillonnage à renouvellement de panel, les données longitudinales sur la séquence (généralement brève) de situations peuvent être obtenues en appariant des données relatives à des individus qui participent à deux ou à plusieurs enquêtes successives. Dans les EPA, la période de référence, ainsi que les notions et les définitions utilisées pour classer les répondants sont généralement conformes aux recommandations du Bureau international du travail (BIT) (Hussmanns, Mehran et Verma 1990); cela donne des mesures des situations d'activité assez précises et comparables entre elles dans l'espace et dans le temps. Les données relatives à l'activité sont recueillies également par l'entremise d'enquêtes-ménages polyvalentes. Dans ces enquêtes, on accorde moins d'attention à la situation par rapport au marché du travail que dans le cas des enquêtes mentionnées précédemment, et les périodes de référence, ainsi que les notions et les définitions peuvent être moins conformes aux recommandations du BIT.

Les données longitudinales peuvent être recueillies également par l'entremise d'enquêtes rétrospectives. Des enquêtes transversales peuvent comprendre des questions rétrospectives qui servent à obtenir des informations sur la séquence de situations d'activité qui sont vécues par les répondants. Dans ce cas, la stratégie d'interrogation a une importance déterminante en ce qui a trait à la réduction des erreurs de mémoire, des télescopes, etc. Les méthodes utilisées pour améliorer l'exactitude des déclarations dans les enquêtes rétrospectives sont fondées sur des éléments de la psychologie cognitive et sur les techniques d'enquête (pour une analyse de ces méthodes, voir O'Muircheartaigh 1996). En outre, de nombreuses études empiriques fournissent des preuves du nombre et de l'orientation de biais dus à des erreurs de mémoire. Il est utile d'ajouter à cet égard que dans le cas d'études rétrospectives, des facteurs liés à la durée de la période pour laquelle le répondant doit faire appel à sa mémoire, l'importance des événements qui sont considérés ou la difficulté que comporte la récupération de données relatives à des événements passés mènent généralement à une simplification des questions non conforme aux conventions du BIT relatives aux enquêtes sur la population active.

La pratique répandue qui consiste à utiliser une combinaison de stratégie par panel et de stratégie rétrospective offre des possibilités intéressantes pour estimer des flux bruts de population active en présence d'erreurs de classification. Les enquêtes par panel font appel à des questions rétrospectives, du moins pour un nombre limité de sujets, afin de couvrir la période comprise entre deux vagues successives (c'est le cas notamment de la Survey of Income and Program Participation, comme nous le verrons dans la section 4.2). Lorsqu'on utilise une telle stratégie mixte, les principales caractéristiques du processus de mesure doivent être considérées attentivement, étant donné que celles-ci peuvent avoir une incidence considérable sur la formulation de modèles sensés permettant de tenir compte des erreurs de

classification. Des caractéristiques plus spécifiques du processus de mesure découlent également de la prise en considération des particularités du plan d'échantillonnage. Dans une optique différente, la disponibilité de mesures multiples de la situation par rapport au marché du travail, c'est-à-dire des données sur la situation d'activité d'une personne à une époque donnée qui sont fournies par deux ou plusieurs sources différentes, offre une excellente possibilité de modéliser des erreurs de classification. Ces informations ont une grande importance, d'une façon générale, et en particulier lorsqu'il faut prendre en considération des types d'erreurs de classification corrélées assez compliquées. Des indicateurs multiples de la situation d'activité d'une personne peuvent être recueillis lors d'une même interview, ou lors d'interviews distinctes (par exemple, lors de vagues différentes d'une enquête par panel). Le premier cas n'est pas très courant, mais parfois, des questions portant sur la situation d'activité sont posées dans des contextes différents et de diverses façons. Par exemple, on demande d'abord une auto-classification du répondant par rapport au marché du travail; puis, dans une autre section du questionnaire, on pose une série de questions qui permettent de classer le répondant d'après des définitions standard relatives à la population active. (Pour un exemple différent, voir le cas de la Survey of Income and Program Participation, dans la section 4.2.)

Le deuxième cas couvre plusieurs situations. Au moins deux d'entre elles méritent d'être étudiées:

- a) la situation où les données proviennent d'études comportant des réinterviews, et qui sont souvent recueillies spécialement pour obtenir des informations sur les probabilités d'erreurs de classification (dans un tel cas, la pratique courante consiste à assimiler les données de réinterview à des données de validation; pour en savoir davantage sur les méthodes classiques utilisées pour corriger des flux bruts d'après des données de réinterview, voir Abowd et Zellner 1985, Poterba et Summers 1986, ainsi que Chua et Fuller 1987);
- b) la situation où les données sont recueillies rétrospectivement dans le cadre d'enquêtes par panel mais renvoient à une époque déjà couverte par l'interview précédente, ou sont recueillies lors d'une enquête supplémentaire qui est effectuée occasionnellement et qui couvre la ou les périodes de référence de l'enquête par panel actuelle. Il est évident que dans ce cas, des mesures différentes d'une ou de plusieurs variables peuvent être entachées d'erreurs de classification comportant en général des caractéristiques différentes.

Un grand nombre de questions qui sont soulevées ici sont expliquées dans les études de cas présentées dans la section 4, où la présence simultanée de données de panel et

précitée certains modèles bien connus qui sont utilisés pour la correction de flux bruts observés (section 3.1). Nous présentons ensuite un cadre pratique pour la formulation des modèles susmentionnés; ce cadre est fourni par des modèles de classe latente, et plus précisément par le modèle «LISREL modifié» qui a été proposé par Hagenaars (1990), et qui est un outil général qui permet de décrire de type causal existant entre des variables nominales observées et non observées (section 3.2).

La dernière et principale partie de l'article (section 4) est consacrée à une présentation détaillée de deux études de cas. Nous appliquons la même méthode de modélisation aux deux cas: des informations a priori sur les caractéristiques de mesure de l'enquête (et si possible, sur le processus réel) sont combinées à des recherches de spécifications, afin d'obtenir des modèles parcimonieux et (nous l'espérons) sensés. Comme nous l'avons mentionné précédemment, ces deux études de cas sont raisonnablement différentes, principalement en ce qui a trait au plan d'échantillonnage; cette différence s'avère utile pour illustrer diverses spécifications de modèle et diverses stratégies utilisées pour obtenir et tester la formulation finale.

Concernant les deux études de cas, on constate ce qui suit:

- a) le modèle LISREL modifié s'est avéré suffisamment souple pour modéliser le mécanisme à l'origine des erreurs contenues dans les données longitudinales qui ont été recueillies d'après des plans d'échantillonnage différents, ainsi que pour modéliser le processus à l'origine des transitions réelles au sein de la population active;
- b) en particulier dans la partie du modèle relative aux mesures, nous avons été en mesure d'intégrer le patron d'erreurs corrélées ainsi que les effets que produisent ces erreurs, qui sont particulièrement importants dans le cas d'enquêtes qui comportent des éléments rétrospectifs;
- c) les transitions observées sont corrigées dans le sens prévu d'après des preuves théoriques et empiriques relatives aux effets des erreurs de mesure (et non automatiquement dans le sens de la mobilité, comme c'est le cas avec des stratégies fondées sur l'hypothèse d'erreurs de classification indépendantes).

2. LE RÔLE DES STRATÉGIES DE COLLECTE DE DONNÉES

Les informations servant à l'estimation des flux bruts relatifs à la population active proviennent de données longitudinales, c'est-à-dire d'observations portant sur les mêmes unités et effectuées à diverses époques. Depuis peu, on s'efforce davantage de recueillir des données longitudinales. Cela est vrai également dans le cas d'enquêtes dont le but principal est d'évaluer la situation d'individus au sein de la population active, dans le cas d'une population

donnée. D'autre part, cet intérêt à l'égard de la collecte et de l'utilisation de données longitudinales soulève de nouvelles questions concernant l'origine et le patron des erreurs de mesure (ou de classification), ainsi qu'au sujet des effets possibles de ces erreurs sur les estimations des quantités qu'on souhaite connaître. Parmi les études de référence générales portant sur les sources d'erreurs de classification touchant des données longitudinales recueillies dans le cadre d'enquêtes au fil du temps, il y a celles de Duncan et Kalton (1987) et de Kalton et Citro (1993). Dans la présente section, nous abordons brièvement certains des principaux effets des erreurs de classification sur les stratégies de modélisation utilisées pour corriger des flux bruts.

Un argument qui est avancé habituellement au sujet de l'incidence des erreurs de mesure dans l'estimation de flux bruts est que ces erreurs mènent à une surestimation des variations. Cela est vrai lorsqu'on présuppose que les erreurs de mesure ne sont pas corrélées au fil du temps. Or, dans de nombreux cas, cette hypothèse n'est pas réaliste (voir Skinner et Torelli 1993; Singh et Rao 1995; van de Pol et Langheine 1997) et devrait être reconsidérée en tenant compte attentivement de la stratégie de collecte de données qui a été effectivement adoptée. De manière générale, si les données longitudinales sont recueillies par interrogation rétrospective (du moins en partie), on peut affirmer que des erreurs de mémoire entraînent des erreurs corrélées.

Des hypothèses précises concernant les erreurs de relatives à la situation au sein de la population active. Les stratégies de modélisation utilisées pour corriger des flux bruts entachés d'erreurs de classification devraient aussi tenir compte du processus de mesure qui est effectivement utilisé, étant donné que le nombre d'erreurs de classification et la direction d'éventuels biais sont liés à la stratégie qui a été adoptée pour recueillir les données longitudinales. Comme on le sait bien, des données longitudinales peuvent être obtenues à l'aide de diverses stratégies d'enquête. Il est utile à cet égard de faire tout au moins la distinction entre, d'une part, les enquêtes par panel et, d'autre part, les enquêtes rétrospectives. En outre, la disponibilité d'indicateurs multiples mérite une attention particulière.

Les enquêtes par panel représentent la façon la plus naturelle de recueillir des données longitudinales. Parmi ce type d'enquêtes, les enquêtes à renouvellement de panel jouent un rôle important. En effet, il s'agit du type de méthode qui est utilisé pour la plupart des enquêtes sur la population active (EPA), dont le but principal est l'estimation des statistiques de la population active. Dans

Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification

FRANCESCA BASSI, NICOLA TORELLI et UGO TRIVELLATO¹

RÉSUMÉ

Les flux bruts entre situations d'activité sont d'une grande importance pour comprendre la dynamique du marché du travail. Les flux observés sont généralement entachés d'erreurs de classification qui peuvent introduire des biais importants. Dans le présent article, nous passons en revue certaines des stratégies les plus couramment utilisées pour recueillir des données longitudinales sur la situation par rapport au marché du travail, ainsi que des méthodes de modélisation qui ont été créées pour corriger les flux bruts lorsqu'ils présentent des erreurs de classification. Nous présentons un cadre général pour l'estimation des flux bruts. Nous donnons des exemples de diverses spécifications de modèle appliquées à des données qui ont été recueillies avec des stratégies différentes. Nous considérons en particulier deux cas, c'est-à-dire les flux bruts de la U.S. Survey of Income and Program Participation et les flux bruts de l'Enquête française sur la population active, qui est une enquête annuelle dans laquelle on recueille des données mensuelles rétrospectives.

MOTS CLÉS : Erreurs de classification corrélées; modèles de classe latente; données longitudinales; erreurs de mémoire; effet de «liste».

1. INTRODUCTION

Les flux bruts entre situations d'activité représentent un puissant moyen pour analyser la dynamique du marché du travail. Les mouvements bruts ont trait à des variations qui ont lieu au niveau des individus, leur estimation dépend donc de la disponibilité de données longitudinales. Les effets d'une classification erronée d'unités par rapport à leur situation au sein de la population active peuvent être à l'origine de fausses transitions. Même si l'on peut présupposer que ces erreurs s'annulent lors de l'estimation de flux nets, on ne peut pas ne pas en tenir compte dans l'estimation des flux bruts. Diverses stratégies permettent de corriger les flux bruts afin de tenir compte d'erreurs de classification. Essentiellement, ces stratégies sont fondées sur

- a) des hypothèses concernant l'origine des erreurs de classification, qui peuvent être dues
a1) au plan d'échantillonnage (enquêtes par panel, éventuellement à renouvellement de panel; enquêtes rétrospectives; des combinaisons d'enquête par panel et d'enquête rétrospective, etc.), ou
a2) au contenu et à la structure du questionnaire (existence d'un ou de plusieurs indicateurs de la variable observée, types de questions (questions fondées sur un épisode ou sur un événement), etc.);
- b) des hypothèses concernant le processus qui est à l'origine des transitions entre les situations relatives à la population active.

Dans le présent article, nous passons en revue certaines des stratégies les plus couramment utilisées pour recueillir des données longitudinales sur la situation d'activité, ainsi que des méthodes de modélisation qui ont été créées pour corriger des flux bruts lorsqu'ils présentent des erreurs de classification. Nous montrons que la plupart des spécifications habituelles qui sont proposées dans les textes spécialisés peuvent être considérées comme des particuliers d'une formulation générale qui permet de déterminer les avantages et les inconvénients de chaque spécification et d'envisager une stratégie d'estimation commune. Le présent article porte principalement sur des applications sensées de cette méthode de modélisation générale dans l'estimation de flux bruts à partir de données d'enquête recueillies par des stratégies différentes. Nous considérons deux cas en particulier: la U.S. Survey of Income and Program Participation et l'Enquête française sur la population active, qui est une enquête annuelle à renouvellement de panel comportant des informations mensuelles rétrospectives. L'organisation de l'article est décrite ci-après. Dans la section 2, nous abordons brièvement diverses stratégies de collecte de données longitudinales sur l'activité, ainsi que le rôle probable que jouent ces stratégies dans l'introduction d'erreurs de classification, d'après leur description dans les études qui traitent des méthodes d'enquête. Dans la section 3, nous présentons une méthode assez générale qui permet de modéliser des flux bruts entachés d'erreurs de classification, c'est-à-dire qui permet d'estimer à la fois des flux bruts réels et des probabilités de réponse conditionnelles. Nous donnons également des exemples de la façon dont on peut spécifier comme cas particuliers de la méthode

¹ Francesca Bassi, Nicola Torelli et Ugo Trivellato, Dipartimento di Scienze Statistiche, Via San Francesco, 33, 35121, Padova, Italy.

- HOLT, D., et SKINNER, C.J. (1989). Components of change in repeated surveys. *Revue Internationale de Statistique*, 57, 1-18.
- HUGGINS, V.J., et FISCHER, D.P. (1994). The redesign of the survey of income and program participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.
- HUGHES, S., et HINKINS, S. (1995). Creation of panel data from cross-sectional surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-413.
- KALTON, G., et BRICK, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-51.
- KALTON, G., et CITRO, C. (1993). Enquêtes par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.
- KASPRZYK, D., DUNCAN, G., KALTON, G., et SINGH, M.P. (Eds.) (1989). *Panel Surveys*. Wiley: New York.
- LAVALLÉE, P. (1995). Pondération transversales des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-36.
- LENGACHER, J.E., SULLIVAN, C.M., COOPER, M.P., et GROVES, R.M. (1995). Once reluctant, always a reluctant? Effects of differential incentives on later survey participation in a longitudinal study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1029-1034.
- McGUGAN, K.A., ELLICKSON, P.L., HAYS, R.D., et BELL, R.M. (1995). Tracking, weighting, and sample selection modeling to correct for attrition. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 402-407.
- MICHAUD, S., DOLSON, D., ADAMS, D., et RENAUD, M. (1995). Combining administrative and survey data to reduce respondent burden in longitudinal surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 11-20.
- MURRAY, T.S., MICHAUD, S., EGAN, M., et LEMAITRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 715-730.
- HOLT, D., et SKINNER, C.J. (1989). Components of change in repeated surveys. *Revue Internationale de Statistique*, 57, 1-18.
- HUGGINS, V.J., et FISCHER, D.P. (1994). The redesign of the survey of income and program participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 668-673.
- HUGHES, S., et HINKINS, S. (1995). Creation of panel data from cross-sectional surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 408-413.
- KALTON, G., et BRICK, J.M. (1995). Méthodes de pondération pour les enquêtes par panel auprès des ménages. *Techniques d'enquête*, 21, 37-51.
- KALTON, G., et CITRO, C. (1993). Enquêtes par panel: ajout d'une quatrième dimension. *Techniques d'enquête*, 19, 217-227.
- KASPRZYK, D., DUNCAN, G., KALTON, G., et SINGH, M.P. (Eds.) (1989). *Panel Surveys*. Wiley: New York.
- LAVALLÉE, P. (1995). Pondération transversales des enquêtes longitudinales menées auprès des individus et des ménages à l'aide de la méthode du partage des poids. *Techniques d'enquête*, 21, 27-36.
- LENGACHER, J.E., SULLIVAN, C.M., COOPER, M.P., et GROVES, R.M. (1995). Once reluctant, always a reluctant? Effects of differential incentives on later survey participation in a longitudinal study. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1029-1034.
- McGUGAN, K.A., ELLICKSON, P.L., HAYS, R.D., et BELL, R.M. (1995). Tracking, weighting, and sample selection modeling to correct for attrition. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 402-407.
- MICHAUD, S., DOLSON, D., ADAMS, D., et RENAUD, M. (1995). Combining administrative and survey data to reduce respondent burden in longitudinal surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 11-20.
- MURRAY, T.S., MICHAUD, S., EGAN, M., et LEMAITRE, G. (1991). Invisible seams? The experience with the Canadian Labour Market Activity Survey. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 715-730.
- RIZZO, L., KALTON, G., et BRICK, M. (1994). Adjusting for panel nonresponse in the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 422-427.
- SCHERBAL, J.A., et LAVARAKAS, P.J. (1995). Panel attrition in a dual-frame local area telephone survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1035-1039.
- SCHUBERT, F., et WINKLER, W.E. (1993). Analyse de régression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.
- SINGH, A.C., et WHITRIDGE, P. (Eds.) (1990). *Analysis of data in time. Recueil: Symposium 89, L'analyse des données dans le temps*. Statistique Canada.
- SINGH, A.C., WU, S., et BOYER, R. (1995). Longitudinal survey nonresponse adjustment by weight calibration for estimation of gross flows. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 396-401.
- SINGH, R.P., PETRONI, R.J., et ALLEN, T.M. (1994). Oversampling in panel surveys. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 674-679.
- STEARNS, S.C., KOVAR, M.G., HAYES, K., et KOCH, G.G. (1996). Estimates of national hospital use from administrative data and personal interviews. *Journal of Official Statistics*, 12, 47-61.
- TAMBAY, J.L., SCHIOPU-KRATINA, I., MAYDA, J., STUKEL, D., et NADON, S. (1998). Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population. *Techniques d'enquête*, 24, 159-169.
- TIN, J. (1996). Program participation and attrition: The empirical evidence. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 669-674.
- WEBBER, M. (1994). The survey of labour and income dynamics: lessons learned in testing. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 85-99.

REMERCIEMENTS

L'auteur tient à remercier les examinateurs et le rédacteur adjoint, pour leurs nombreuses suggestions utiles.

BIBLIOGRAPHIE

- ALLEN, T.M., et PETRONI, R.J. (1994). Mover nonresponse adjustment research for the Survey of Income and Program Participation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 662-667.
- AN, A.B., BREIDT, F.J., et FULLER, W.A. (1994). Regression weighting methods for SIPP data. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 434-439.
- ARMSTRONG, J., DARCOVICH, N., et LAVALLÉE, P. (Éds.) (1993). *Recueil: Symposium 92, Conception et analyse des enquêtes longitudinales*, Statistique Canada.
- BASSI, F., TORELLI, N., et TRIVELLATO, U. (1998). Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification. *Techniques d'enquête*, 24, 117-132.
- BINDER, D.A., et HIDROGLOU, M.A. (1988). Sampling in time. *Handbook of Statistics Volume 6: Sampling* (Éds. P.R. Krishniah et C.R. Rao). North Holland: Amsterdam, 187-211.
- CLARKE, P.S. et CHAMBERS, R.L. (1998). Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage. *Techniques d'enquête*, 24, 133-140.
- CORDER, L.S., MANTON, K.G., et WOODBURY, M. (1994). Improving coverage, response rates, and nonresponse follow-up via a longitudinal list sample design: The National Long-Term Care Surveys. *Proceedings of the 1994 Annual Research Conference*, U.S. Bureau of the Census, 63-84.
- CZAJKA, J.L. (1994). Income stratification in panel surveys: Issues in design and estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 791-796.
- DORINSKI, S.M., et HUANG, H. (1994). Use of administrative data in SIPP longitudinal estimation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 656-661.
- DUNGAN, G.J., et KALTON, G. (1987). Issues of design and analysis of surveys across time. *Revue Internationale de Statistique*, 55, 97-117.
- FOLSOM, R.E., et WITT, M.B. (1994). Testing a new attrition nonresponse adjustment method for SIPP. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 428-433.
- HIDROGLOU, M.A., et MICHAUD, S. (Éds.) (1998). *Recueil: Symposium 98, Analyses longitudinales pour enquêtes complexes*, Statistique Canada. À paraître.
- HILL, D.H. (1994). The relative empirical validity of dependent and independent data collection in a panel survey. *Journal of Official Statistics*, 10, 359-380.

siques utilisés pour l'analyse des données dans le temps ne comportent pas de poids d'échantillonnage. Il faut donc mettre au point des méthodes pour intégrer efficacement les poids d'échantillonnage dans l'analyse. L'utilisation de poids s'avère souvent la méthode à privilégier dans le cas des enquêtes à grande échelle, car ceci aide à éviter l'erreur de spécification du modèle.

Dans certains cas, les erreurs dues au traitement des données, par exemple au couplage des enregistrements, devront être incluses dans l'analyse ou, du moins, faire l'objet d'une étude pour en comprendre l'incidence (voir, par exemple, Dorinski et Huang 1994).

Les données administratives sont souvent utilisées pour l'analyse, car ces données sont parfois plus facilement utilisables que les données recueillies. Cependant, comme les données administratives peuvent présenter des problèmes au niveau conceptuel ou autre, des méthodes d'analyse spéciales sont parfois requises pour assurer un usage efficace de ces données.

Enfin, il y a lieu de mentionner les difficultés associées à la diffusion des données. Pour de nombreux phénomènes, il faut défrayer des mesures longitudinales sommaires. Or, souvent, ces mesures ne cadrent pas avec les tableaux habituellement utilisés pour les enquêtes transversales. De plus, un grand nombre d'analyses requièrent un accès à des micro-données ce qui, en retour, risque de créer des problèmes en ce qui a trait à la protection de la confidentialité de l'information. En effet, il se peut que les mesures habituelles prises au moment de la diffusion de fichiers de micro-données provenant d'enquêtes transversales ne suffisent pour la diffusion des résultats d'enquêtes de nature longitudinale, du fait que les bases de données longitudinales sont beaucoup plus riches et donc que la possibilité de pouvoir identifier une personne à partir de ces données devient beaucoup plus grande. Comme la protection de la confidentialité du répondant est d'une importance capitale, une approche conservatrice doit parfois être utilisée, même si celle-ci ne répond pas entièrement aux exigences des utilisateurs.

SOMMAIRE

Nous avons discuté brièvement d'un grand nombre de questions auxquelles s'intéressent actuellement les chercheurs qui étudient la conception et l'analyse des études longitudinales et nous avons constaté que plusieurs d'entre elles devront faire l'objet d'une analyse plus poussée. Des réponses à bon nombre de ces questions apparaîtront à mesure que notre expérience des enquêtes longitudinales augmentera; cependant, cet approfondissement de nos connaissances fera naître également beaucoup de nouvelles questions. Ceci ouvre donc la voie à de nombreuses études et recherches significatives.

par des méthodes économétriques). Pour leur part, Schejbál et Lavrakas (1995) étudient l'effet de l'attrition du panel lors d'une enquête téléphonique locale à double base de sondage. Corder, Manton et Woodbury (1994) étudient des méthodes pour améliorer la couverture et réduire l'attrition dans le cadre d'une enquête nationale sur les soins de longue durée (National Long Term Care Survey). L'attrition du panel peut être due à un refus de participer ou à l'impossibilité de rejoindre les personnes; or l'effet de l'attrition peut être très différent selon qu'il s'agit d'une enquête longitudinale ou transversale, et ces différences doivent être étudiées. Allen et Petroni (1994) discutent quant à eux du problème de la correction des données en fonction des personnes qui ont déménagé.

Enfin, il faut concevoir des études de contrôle de la qualité qui tiennent compte des caractéristiques particulières des enquêtes longitudinales. Un grand nombre d'études sur le contrôle de la qualité, outre les études habituelles prévues pour les enquêtes transversales, peuvent être utilisées pour les enquêtes longitudinales, car la répétition de l'enquête peut permettre une identification plus efficace des cas sujets à erreur. Comme la stabilité des données dans le temps est importante pour les études longitudinales, les changements méthodologiques peuvent avoir un impact sur les mesures longitudinales à l'étude et ceux-ci doivent être étudiés. Enfin, les données administratives peuvent fournir des évaluations utiles, certaines de ces données pouvant aider à valider les résultats.

QUESTION 4: ANALYSE

Le dernier volet de notre examen porte sur les questions liées à l'analyse, et principalement sur l'analyse potentielle de l'étude longitudinale. Les causes ou déterminants des divers résultats présentent un grand intérêt pour les utilisateurs de données. Cependant, la modélisation de ces causes peut être complexe, en particulier si l'enquête elle-même est complexe. Un grand nombre de ces questions sont examinées dans Singh et Whitridge (1990) et dans Hidiroglou et Michaud (1998).

Les mesures des flux bruts et autres mesures du changement brut sont quelques exemples des types d'analyses courantes. Flux brut fait référence au changement, pour une personne, d'une catégorie à une autre. Il s'agit, en d'autres mots, de la transition de la catégorie *A* à la catégorie *B* entre deux moments précis, alors que le flux net correspond au résultat net du changement pour un ensemble de personnes, dans le temps. L'incidence d'une erreur dans la mesure du flux brut soulève des questions difficiles. S'il se produit des erreurs de mesure assez importantes à chaque cycle, il en résultera un impact significatif sur le biais des estimations du flux brut, même si cela n'a aucune incidence sur le flux net en soi. Parfois, le renouvellement de l'échantillon aggrave ce problème, car il peut être difficile de tenir compte adéquatement du renouvellement de

l'échantillon lorsqu'on mesure les flux bruts. Il faut prévoir un traitement spécial pour les panels qui s'ajoutent à l'échantillon à un cycle donné et pour ceux qui ont été éliminés de l'échantillon au cycle précédent, afin d'obtenir une bonne estimation de ces flux. Pour la mesure des flux bruts, il importe également de différencier les changements dans la population des flux bruts proprement dits. Autrement dit, les changements qui surviennent d'un cycle à un autre sont une combinaison des changements dans la taille de la population et des changements à l'intérieur de la population. La situation peut devenir encore plus complexe lorsque les flux bruts sont eux-mêmes analysés en regard d'autres données, par exemple la dynamique du revenu.

Comme l'a fait remarquer un examinateur, il est important d'éduquer les utilisateurs sur la façon d'analyser efficacement les données longitudinales. La récente prolifération des enquêtes longitudinales fait naître de nombreuses possibilités en termes d'analyses nouvelles; cependant, il est possible que bon nombre des analyses qui n'ont été étudiées jusqu'ici que les enquêtes transversales ne connaissent pas les techniques qui conviennent le mieux aux enquêtes longitudinales.

Pour les nombreuses enquêtes qui utilisent des bases de sondage s'appuyant sur des données administratives, il peut s'avérer nécessaire de tenir compte des changements dans la base de sondage durant l'analyse, car les inclusions dans la base de sondage peuvent dépendre des changements apportés aux procédures administratives et des changements dans les conditions propres à la personne. La composition d'un fichier des prestataires de l'assurance-chômage, par exemple, évolue en fonction des critères d'admissibilité et de la situation personnelle de l'individu.

La mesure du changement peut souvent être décomposée en diverses composantes. À titre d'exemple, une distinction peut être établie entre les mouvements des unités de l'échantillon, d'un domaine à un autre, et les changements dans les données, pour les unités d'un même domaine. Holt et Skinner (1989) présentent une discussion intéressante sur diverses composantes du changement.

Dans le cas des analyses plus complexes, comme la modélisation des séries chronologiques, la plupart des modèles classiques de séries chronologiques ne prennent pas en considération le fait que l'information provient d'un sondage. Par conséquent, la modélisation des séries chronologiques ne tient pas compte adéquatement des erreurs d'échantillonnage attribuables au type d'enquête.

Par ailleurs, certaines mesures peuvent dépendre d'autres enquêtes transversales. Il se peut, par exemple, que les tranches de revenu utilisées pour l'analyse de l'enquête longitudinale soient tirées d'une autre enquête, celle-ci de nature transversale; or ceci risque d'accroître la complexité de l'analyse, car ces tranches ne sont pas nécessairement stables.

L'utilisation ou non de poids d'échantillonnage et la façon de les utiliser, le cas échéant, créent des difficultés pour nombre d'analystes, car de nombreux modèles clas-

l'enquête. Il peut y avoir des avantages à utiliser l'imputation pour des études longitudinales plutôt que transversales. Dans le premier cas, l'imputation est basée sur les données longitudinales contenues dans la base de données pour le même répondant, et non sur l'information obtenue des autres répondants, au même moment. Pour résoudre les problèmes d'attrition et de non-réponse cyclique, on peut modéliser les taux d'attrition et utiliser ces modèles pour compenser pour la non-réponse, par ajustement des poids. Diverses méthodes d'ajustement des poids ont été examinées pour la Survey of Income and Program Participation et les résultats sont présentés dans Rizzo, Kalton et Brick (1994), Folsom et Witt (1994) et An, Bredt et Fuller (1994). Singh, Wu et Boyer (1995) ont pour leur part étudié ce problème, en regard du cas difficile que présente l'estimation des flux bruts.

Le calcul des poids peut poser de nombreuses difficultés. Il existe diverses approches et techniques pour calculer les poids transversaux et longitudinaux. Les premiers sont utilisés pour les mesures de la population qui font référence à un moment précis dans le temps, alors que les deuxièmes doivent être utilisées lorsque sont incluses des données sur plus d'un cycle. L'analyste peut choisir d'utiliser des poids propres à la personne, qui diffèrent des poids qui s'appliquent à des ménages (voir Kalton et Brick 1995). À titre d'exemple, pour certaines variables comme le revenu familial, il serait préférable d'utiliser des poids qui s'appliquent aux ménages plutôt qu'à la personne. La pondération devient plus complexe lorsqu'il y a utilisation de bases de sondage multiples. L'utilisation efficace de données administratives peut engendrer des complications encore plus grandes dans le plan de pondération lui-même (voir, par exemple, Stearns et coll. 1996).

De nombreux facteurs peuvent rendre un échantillon sous-représentatif. Ainsi, l'absence de représentativité peut être due à des problèmes de couverture résultant d'une immigration dans la population. Dans d'autres cas, le sous-dénombrement peut être attribuable à l'attrition. À l'opposé, le surdénombrement peut résulter de l'inclusion de certains cohabitants non échantillonnés d'un ménage, ce qui suppose que ces personnes sont incluses dans l'échantillon du seul fait qu'elles vivent avec une personne appartenant à l'échantillon initial (voir Lavallée 1995 et Kalton and Brick 1995). D'autres types de surdénombrement général peuvent se produire et des techniques de pondération spéciales doivent être utilisées pour s'assurer qu'aucun biais n'est introduit. Cependant, dans le cas notamment des enquêtes longitudinales, ceci peut devenir assez complexe. Les données administratives peuvent alors être utilisées pour déterminer si, oui ou non, l'échantillon est représentatif et pour fournir de l'information en vue d'apporter les corrections nécessaires.

Comme la majeure partie des estimations, aux fins d'une étude longitudinale, consistent à mesurer des changements plutôt qu'à mesurer un phénomène à un moment précis, ceci soulève des questions quant à la façon de définir les

QUESTION 3: ÉVALUATION

Le troisième aspect dont nous discutons ici est celui de l'évaluation des données et des méthodes. Même si les évaluations peuvent se faire indépendamment de la mise en oeuvre, les résultats de ces évaluations devraient influencer l'enquête elle-même, soit en modifiant les méthodes d'estimation, soit en modifiant la façon dont l'enquête est conçue et mise en oeuvre durant les cycles subséquents.

Il existe de nombreuses sources de biais qui peuvent être étudiées. Le biais peut être dû à l'utilisation d'une technique d'interview dépendante où le répondant et l'intervieweur reçoivent de l'information faisant référence à un cycle antérieur de l'enquête. L'effet de liste peut résulter d'études rétrospectives; voir, par exemple Murray, Michaud, Egan et Lemaître (1991). D'autres sources de biais peuvent être introduites, lorsque la non-réponse est informative, c'est-à-dire lorsque la propension à la non-réponse est liée à la variable à l'étude, par exemple lorsque la non-réponse dans le ménage est liée aux flux bruts à l'intérieur du ménage (flux bruts faisant ici référence aux changements dans la classification de la personne); voir Clarke et Chambers (1989). D'autres biais peuvent être dus à des erreurs de mesure ou de classification: voir, par exemple, Bassi, Torelli et Trivellato (1998). Enfin, il peut y avoir introduction d'un biais de conditionnement du fait que le répondant devient plus sensible à certaines questions qui lui sont posées, par exemple sur la dynamique du travail, et donc que son comportement change du seul fait de sa participation à l'enquête.

Il convient également d'évaluer l'effet des erreurs de réponse et des erreurs de l'intervieweur sur l'analyse. À des méthodes d'interview différentes peuvent correspondre des taux d'erreur différents. Le fait qu'il y ait ou non roulement des intervieweurs peut aussi avoir une incidence sur certaines analyses, tout comme l'utilisation ou non de répondants substitués.

D'autres évaluations peuvent être faites pour mesurer l'effet de l'attrition et évaluer diverses techniques d'imputation et autres méthodes de traitement de la non-réponse (voir Tin 1996, pour une évaluation de l'attrition

étape consiste à choisir le mode de collecte des données, parmi les divers choix qui s'offrent. L'interview assistée par ordinateur est une technique qui se répand de plus en plus et qui élargit le choix des instruments d'enquête pouvant être utilisés. Cette technique facilite entre autres l'utilisation de l'interview dépendante, où le répondant ou l'intervieweur a accès aux réponses des cycles précédents; ceci peut augmenter, ou réduire, certains biais. Hill (1994) évalue cette méthode en regard de la Survey of Income and Program Participation aux États-Unis.

Bien sûr, comme ce sont les mêmes répondants qui sont interrogés à diverses reprises, la question du fardeau de réponse prend alors encore plus d'importance que lors d'une enquête transversale unique. Il faut éviter de surcharger le répondant, car ceci risque d'accroître le taux de refus aux cycles subséquents. Michaud, Dolson, Adams et Renaud (1995) proposent de réduire le fardeau du répondant en faisant un usage accru des données administratives. La réduction de l'attrition due à la non-réponse est un objectif important des enquêtes longitudinales et il y aurait lieu d'envisager l'utilisation d'incitatifs financiers ou autres pour aider à préserver l'intégrité de l'échantillon au fil des ans (voir Lengacher, Sullivan, Couper et Groves 1995). Un autre moyen de réduire l'attrition est de recueillir de l'information pour faciliter les efforts de recherche des répondants et aussi de garder contact avec les répondants au fil des ans; McGuigan, Ellickson, Hays et Bell (1995) examinent diverses méthodes de recherche, de pondération et de modélisation de la sélection de l'échantillon, pour résoudre les problèmes d'attrition.

Dans certaines enquêtes longitudinales, la collecte des données se fait de façon rétrospective, c'est-à-dire que les questions posées font référence à la fois à des événements antérieurs et présents; une telle situation peut donner lieu à un effet dit de hystérie. Les changements observés durant les périodes de référence peuvent donc dépendre des périodes pour lesquelles les données ont été obtenues de façon rétrospective.

Les dossiers administratifs peuvent être utiles pour enrichir la base de données et éviter ainsi que toutes les données soient recueillies directement du répondant (voir Michaud et coll. 1995). Bien sûr, ceci peut dépendre de la qualité des données administratives, de leur disponibilité et de la réciprocité entre l'information contenue dans les dossiers administratifs et les variables à l'étude; voir Stearns, Kovar, Hayes et Koch (1996) pour un exemple sur cette question. Cependant, lorsqu'on utilise des données administratives ou des fichiers combinés, ces divers fichiers peuvent comporter des lacunes statistiques et il faut alors se demander comment pallier ce problème.

En général, la modification de la structure de la base de sondage peut causer des problèmes durant l'exécution des analyses longitudinales. Des changements peuvent également survenir en ce qui a trait à certaines caractéristiques clés du répondant; dans le cas par exemple d'un registre des entreprises, si l'information sur la classification

de l'industrie change, du fait que les entreprises ont modifié la nature des produits offerts, il est alors important que la base de données reflète ce changement de la classification pour assurer une utilité maximale des analyses longitudinales. Cependant, ceci risque également de compliquer l'analyse.

De nombreux problèmes surviennent, lorsque la base de données est créée en combinant des échantillons prélevés d'une série d'enquêtes individuelles. L'intégration de cette information peut créer des difficultés, du fait que des méthodes différentes ont pu être utilisées pour les diverses enquêtes, ce qui risque de se solder par une variation dans la qualité de l'information, d'une base de données à une autre.

Pour bon nombre d'enquêtes longitudinales, les questions liées au couplage des enregistrements sont importantes. Le couplage est en effet utilisé à différentes étapes du traitement et, dans certains cas, les études longitudinales sont basées uniquement sur ces fichiers couplés. Le couplage des enregistrements est ainsi fréquemment utilisé pour créer et mettre à jour les bases de sondage, entre autres pour lier des fichiers administratifs dans le temps, lier des fichiers administratifs et des bases de sondage ou encore lier des bases de sondage distinctes. Dans le cas par exemple d'enquêtes sur les établissements, il se peut que l'on veuille créer des dossiers longitudinaux composites qui soient basés sur plusieurs enquêtes indépendantes à passages répétées, étant donné que bon nombre des établissements sont interrogés à chaque cycle de l'enquête. Le couplage est souvent utilisé pour déterminer quelles unités correspondent aux mêmes établissements, ainsi que pour déterminer les ajouts à une base de sondage. Bien sûr, les erreurs dues au couplage peuvent être importantes; voir Scheuren et Winkler (1993) à ce sujet.

Dans d'autres cas, aucun répondant véritable n'est suivi au fil des ans. Le couplage des enregistrements sert ici plutôt à créer des populations artificielles, par le biais de l'appariement statistique. Ces populations sont ensuite analysées, comme si elles étaient réelles.

Une autre question liée à la mise en oeuvre est celle du traitement de la non-réponse. On sait que la non-réponse aux enquêtes longitudinales n'est pas entièrement le fruit du hasard et qu'elle tend à varier, selon la sous-population. Une attention particulière doit donc être portée au choix de la méthode d'imputation ou de pondération à utiliser; voir, par exemple, Tambay, Schiopu-Kratina, Mayda, Stukel et Nadon (1998). Lorsque des données administratives servent de fondement à une étude longitudinale, il risque d'y avoir des données manquantes, d'où la nécessité de prévoir des méthodes spéciales pour y remédier.

L'imputation et la pondération sont les deux méthodes habituellement utilisées pour pallier les données manquantes. La pondération est utilisée habituellement lorsque la non-réponse est propre à un cycle particulier. L'imputation est utilisée plus fréquemment lorsqu'on observe une non-réponse partielle durant un cycle donné de

doit assurer le prélèvement d'un échantillon suffisant auprès de la population à l'étude, ainsi que de tout groupe témoin. Les données administratives se sont avérées très utiles pour la conception d'un échantillon pour bon nombre de ces enquêtes, car elles procurent souvent une base de sondage adéquate.

Comme l'a fait remarquer un des examinateurs, un élément clé de la phase de conception est celui de la stratégie à utiliser pour tenir compte de la réduction de l'échantillon sous l'effet de l'attrition, de la non-réponse, du retrait de la population-cible, etc. Au nombre des possibilités qui s'offrent, mentionnons l'élargissement de l'échantillon lors des cycles subséquents; une telle stratégie risque toutefois d'altérer la représentativité de la cohorte. Une autre stratégie serait de commencer avec un échantillon plus large et de ne pas remplacer les unités perdues, cette dernière stratégie étant proposée notamment par Singh, Petroni et Allen (1994).

Avant de convenir d'un plan d'échantillonnage partiel-culier, il faut tenir compte des techniques de pondération et d'estimation connexes. Il faut aussi déterminer la période ou la fréquence de l'enquête. De toute évidence, lorsque les variables à l'étude changent souvent, il serait plus souhaitable de mener l'enquête plus fréquemment. D'une part, des enquêtes plus fréquentes augmentent les coûts et le fardeau de réponse; d'autre part, une fréquence réduite peut accroître le biais dû à l'erreur de mémoire. Ces compromis entre les coûts et la qualité sont habituellement difficiles à quantifier.

Très souvent, lorsqu'on a besoin à la fois d'estimations transversales et longitudinales, il faut s'assurer d'avoir des estimations transversales valides pour la sélection des échantillons supplémentaires, ceci du fait qu'il peut y avoir des membres de la population dans les estimations transversales qui ne faisaient pas partie de la base de sondage lors des cycles précédents et qui ne seraient donc pas représentés dans l'échantillon. Czajka (1994) étudie ceci en regard de l'estimation du revenu.

Il est également utile, durant la phase de planification, de prévoir des échantillons pour fins d'évaluation. En effet, divers types de biais peuvent être introduits dans les enquêtes longitudinales, dont certains peuvent survenir du seul fait que la collecte des données, à chaque cycle, se fait auprès des mêmes répondants. Aussi, serait-il bon d'envisager l'addition d'échantillons supplémentaires qui serviraient uniquement à des évaluations visant à mesurer certains de ces effets. Ces échantillons seraient composés d'individus de la population-cible qui ne faisaient pas partie de l'enquête longitudinale. De tels échantillons sont des plus utiles pour évaluer les mesures transversales.

QUESTION 2: MISE EN OEUVRE

La deuxième question que nous abordons ici est celle de la mise en oeuvre d'une étude longitudinale. La première

de sondage. Une pratique courante consiste à prendre un certain nombre de fichiers administratifs différents et à les apparier pour créer une base de sondage. D'autres études longitudinales sont basées uniquement sur l'information contenue dans divers fichiers administratifs. Manifestement, la difficulté tient au fait que ces fichiers administratifs évoluent au fil des ans et donc qu'il devient nécessaire de modifier les échantillons prélevés de ces fichiers et aussi d'adopter des mesures spéciales pour s'assurer que les analyses demeurent pertinentes.

L'étude longitudinale est souvent basée sur une enquête existante ou sur un recensement mené antérieurement, qui sert ensuite de fondement à la base de sondage utilisée pour le suivi des répondants au fil des ans. Un des inconvénients d'une telle approche est qu'il devient difficile d'obtenir des estimations transversales lorsque les additions dans la population sont exclues de la base de sondage. Il faut parfois avoir recours à des techniques de couplage des enregistrements pour la mise à jour de la base de sondage – une technique habituellement sujette à l'erreur.

Dans de rares cas, il est souvent avantageux d'utiliser non seulement une, mais plusieurs bases de sondage. Une telle méthode assure une représentation adéquate des populations susceptibles d'être sous-représentées dans une base de sondage unique; cependant, elle peut également nécessiter l'utilisation d'une méthode de couplage des enregistrements et de techniques de pondération complexes. La méthode d'échantillonnage à utiliser, une fois la base de sondage constituée, est un autre aspect important à considérer durant la phase de conception. Kalton et Citro (1993) traitent d'un certain nombre d'enquêtes longitudinales différentes, notamment des enquêtes à passages répétés, c'est-à-dire d'une série d'enquêtes transversales; des enquêtes par panel, où certains répondants sont sélectionnés et suivis dans le temps; des enquêtes par panel à passages répétés, pour lesquelles de nouveaux panels sont sélectionnés à différents moments, durant une période donnée; des enquêtes par panel avec renouvellement, où il y a, à chaque cycle, suppression d'un panel et addition d'un nouveau panel; des enquêtes avec échantillons chevauchants, pour lesquelles de mêmes répondants se retrouvent d'un cycle à un autre sans qu'il s'agisse pour autant d'un plan d'échantillonnage par panel fixe ainsi que des enquêtes par panel fractionné, où il peut y avoir combinaison d'enquêtes par panel et d'enquêtes à passages répétés ou avec renouvellement. Le plan d'échantillonnage

dimensions à leur conception et leur analyse. Dans les quatre sections qui suivent, nous résumons ces questions en regard de quatre volets distincts, soit : la conception, la mise en oeuvre, l'évaluation et l'analyse. Bon nombre de ces questions ont été discutées dans Kasprzyk, Duncan, Kalton et Singh (1989) et Armstrong, Darcovich et Lavallée (1993). Certaines questions liées à la conception et aux séries chronologiques sont traitées dans Binder et Hidiroglou (1988) ; quelques références plus récentes sont également citées.

QUESTION 1: CONCEPTION

La planification préliminaire, durant l'étape de conception, est essentielle au succès de l'étude longitudinale. Il faut en effet s'assurer que seules des informations pertinentes et précises seront recueillies auprès des répondants, afin de maximiser les avantages potentiels de l'étude longitudinale. Ceci suppose qu'il faut planifier dès le départ les analyses qui seront effectuées à partir de l'enquête longitudinale, afin d'assurer la collecte des données nécessaires à ces analyses. Duncan et Kalton (1987) présentent un excellent résumé de bon nombre de ces questions. Webber (1994), pour sa part, décrit la stratégie d'essai utilisée pour la planification de la Survey on Labour and Income Dynamics aux États-Unis. Toujours aux États-Unis, Hugins et Fischer (1994) discutent des plans en vue du remaniement de la Survey of Income and Program Participation, en se basant sur leurs propres expériences. Les études longitudinales peuvent être plus coûteuses qu'une série d'études transversales. Aussi les avantages découlant de la collecte de données longitudinales doivent-ils être encore plus grands, les coûts étant eux-mêmes plus élevés. Il est en outre essentiel de s'assurer du maintien du financement de l'étude longitudinale, car il faut parfois attendre le deuxième ou le troisième cycle de l'enquête avant d'obtenir des résultats concrets. Bien sûr, on ne peut planifier de la même façon une étude longitudinale et une série d'enquêtes transversales dont les données seront combinées pour créer une base de données longitudinales. De toute évidence, la première est préférable; cependant, il arrive souvent, en raison de la structure de l'organisme de sondage, que l'on possède déjà une série de données transversales et donc que le regroupement de ces données offre une solution de rechange acceptable (voir Hughes et Hinkins 1995).

En général, une attention particulière doit être portée à la conception de la base de données, pour toute enquête longitudinale prévoyant des analyses basées sur des mesures longitudinales, comme l'étude d'épisodes ou de périodes. Certains organismes statistiques sont actuellement à faire la transition, des enquêtes transversales aux enquêtes longitudinales. Une telle transition nécessite cependant une bonne planification. À titre d'exemple, la mise à jour des bases de données servant aux enquêtes longitudinales

diffère considérablement des méthodes de mise à jour utilisées pour les enquêtes transversales et il est à prévoir que la conduite d'un nombre de plus en plus grand d'enquêtes longitudinales mettra en lumière de nombreuses questions liées aux infrastructures et à l'organisation, en particulier en ce qui a trait à la mise à jour des bases de données et aux opérations. Ces changements peuvent en outre avoir une grande incidence sur l'organisme statistique.

Durant la planification d'une enquête longitudinale, il est important de déterminer si les utilisateurs auront ou non besoin aussi d'estimations transversales. Veut-on obtenir des renseignements sur les répondants qui participent à l'enquête durant une période donnée et aussi être en mesure de produire des estimations à un moment précis, comme s'il s'agissait d'une enquête transversale? Le cas échéant, il faudra en tenir compte dans la conception et la mise en oeuvre de l'enquête (voir Lavallée 1995). Une telle préoccupation doit également être prise en considération, si les variables à l'étude requièrent la comparaison d'estimations transversales dans le temps, plutôt que de véritables mesures longitudinales comme l'étude des autocorrélations entre des unités communes dans le cadre d'une enquête auprès des entreprises.

Les concepts et les définitions qui sous-tendent les enquêtes longitudinales sont habituellement établis de concert avec les utilisateurs des données. Même la définition de l'unité longitudinale à observer dans le temps peut devoir être clarifiée pour des populations dynamiques, par exemple pour les enquêtes auprès des ménages ou des entreprises. Il est important de comprendre les exigences de l'utilisateur et de discuter de ce qui peut être mesuré dans le temps, tout en maintenant un niveau de qualité adéquat. Durant la planification de l'enquête, ces exigences doivent être examinées avec soin en regard de ce qui est faisable, au plan opérationnel, durant la réalisation de l'enquête proprement dite. Étant donné les coûts potentiels de ces études, il est souvent avantageux de mener au préalable un essai complet, en particulier des questionnaires. Une autre question qui mérite d'être étudiée plus à fond est celle de l'établissement d'un plus grand nombre de mesures longitudinales types, communes entre les pays, ce qui permettrait aux gouvernements et aux chercheurs d'établir de meilleures comparaisons internationales.

Un autre volet important de la conception des études longitudinales concerne la création, l'utilisation et la mise à jour des bases de sondage au fil des ans, de manière à faciliter la mise en oeuvre de l'étude. À titre d'exemple, il peut arriver qu'une enquête par panel sur les établissements soit basée sur un registre des entreprises dont la composition évolue sans cesse sous l'effet des additions, des retraites et des fusions. Dans de telles circonstances, il est important de bien définir les unités à inclure dans ces panels, au fil des ans.

La fréquence accrue des enquêtes longitudinales au cours des dernières années s'explique notamment par le

d'études, comme les programmes de formation professionnelle ou de formation des adultes.

Dans le domaine de la justice et de la victimisation, il existe de nombreux exemples où l'observation des mêmes personnes dans le temps peut être bénéfique. Que l'on pense par exemple au suivi des personnes qui ont été victimes d'un acte de violence, dans le but d'évaluer les répercussions de cet événement à long terme. De même, il peut être utile d'observer les personnes qui ont eu des démêlés avec la justice, pour évaluer leurs profils de comportement subséquents et les déterminants de ces profils.

Les études sur les comportements des consommateurs présentent un grand intérêt pour les commerçants et autres intervenants, notamment celles portant sur les habitudes d'achat des consommateurs. Bon nombre de chercheurs trouveraient ainsi fort utile d'avoir un historique des achats des consommateurs.

Pour les décideurs, les études sur les effets de la fluctuation des paiements de transfert aux particuliers, au fil des ans, peuvent s'avérer très utiles. Or, une étude longitudinale peut déterminer notamment pendant combien de temps la personne dépendra de ces paiements, si certains de ces paiements créent ou non une dépendance, quelles sont les caractéristiques des bénéficiaires et quels sont les effets à long terme d'une participation à divers programmes d'aide.

En matière d'économie, les caractéristiques longitudinales de diverses entreprises présentent un grand intérêt. Elles permettent entre autres de mesurer l'efficacité d'une entreprise, son degré d'utilisation de la technologie, les effets à long terme de cette utilisation et l'évolution de la productivité au fil des ans. Diverses questions éclairantes sur les caractéristiques démographiques des entreprises pourraient être posées, par exemple quelles sont les caractéristiques qui font qu'une entreprise est vouée à l'échec ou quelles sont les conditions économiques favorables à la création d'entreprises. Il est également utile d'examiner les conditions propices à la fusion des entreprises. Ce sont là tous des phénomènes dont la mesure peut être facilitée par des études longitudinales.

Au cours des dernières années, de nombreuses entreprises ont subi des changements structurels et seules les études longitudinales permettent d'observer certains de ces changements à un micro-niveau. En effet, un grand nombre de mesures ne peuvent être estimées que lorsque les répondants sont interrogés plus d'une fois.

L'agriculture est un autre domaine qui suscite de l'intérêt, en raison de l'évolution constante des pratiques et du profil agricoles. Il est ainsi pertinent d'examiner l'évolution des exploitations agricoles, tant en termes de produits cultivés que de la taille des entreprises, ou encore de savoir dans quelle mesure les caractéristiques des exploitants agricoles ont changé.

Comme nous venons de l'indiquer, les études longitudinales offrent de multiples applications et comportent de nombreuses facettes. Il existe également de nombreuses

principales à retenir est le suivant: lorsqu'on cherche à comprendre certains phénomènes dans le temps, il devient alors nécessaire de procéder à une collecte de données de

façons répétées.

Un certain nombre de facteurs liés aux coûts doivent être considérés avant de décider de la nature d'une nouvelle étude longitudinale. Bien sûr, il faut évaluer les avantages en regard de ces divers coûts. Les enquêtes longitudinales offrent la possibilité d'étudier des questions dans de multiples domaines, dont quelques-uns sont examinés ci-après. Dans le domaine des soins de santé, par exemple, il est notamment intéressant d'étudier l'évolution de l'état de santé et les déterminants à l'origine de ces changements. En

d'autres mots, quels sont les facteurs de risque et quel est, en fait, l'incidence de ces risques sur l'état de santé à long terme? En recueillant des données auprès des mêmes personnes sur une période de temps donnée, il devient possible d'évaluer ces facteurs, non seulement dans le cadre d'études à petite échelle, typiques des essais cliniques, mais également d'enquêtes nationales à grande échelle sur la santé de la population. Il faut reconnaître toutefois que le type d'information recueilli dans le cadre d'une enquête longitudinale nationale sera fort différent de l'information obtenue durant un essai clinique.

Un autre domaine où il est intéressant d'obtenir des observations dans le temps est celui du travail et du revenu. Il ne suffit pas, par exemple, de connaître le changement net, dans le temps, de la situation d'activité ou du taux d'activité de la main-d'œuvre. Il est bon également de savoir quelles personnes sont passées, disons de la situation de sans emploi à celle de personne active ou inactive. Les profils d'emploi ont changé ces dernières années. On retrouve aujourd'hui davantage de femmes sur le marché du travail et aussi davantage d'emplois à temps partiel. Les changements d'emploi sont également plus fréquents. Les enquêtes longitudinales peuvent fournir des réponses à bon nombre de questions qui permettent en retour de mieux comprendre ces phénomènes. Il peut être intéressant, par exemple, d'étudier les caractéristiques des postes de débutants occupés par les personnes auparavant en chômage, ou encore l'efficacité de différentes stratégies de recherche d'emplois ou des programmes gouvernementaux de formation.

La durée des épisodes de pauvreté est une autre question qui retient de plus en plus l'attention afin, par exemple, de déterminer combien de temps une personne à faible revenu demeure dans cette situation. Quels sont les facteurs qui déterminent si cette situation sera à long terme? Quelle est l'importance de la scolarité et d'autres facteurs en ce qui a trait à la pauvreté et à la durée des épisodes de pauvreté? Dans le domaine de l'éducation, maintenant, il est intéressant d'examiner la transition de l'école au milieu du travail, au moment de l'entrée sur le marché du travail, et l'étude longitudinale peut convenir davantage que les autres études à l'examen de ce phénomène. Un autre exemple relie aux études est celui de l'efficacité des divers programmes

Les enquêtes longitudinales: Pourquoi ces enquêtes diffèrent-elles de toutes les autres enquêtes?

DAVID A. BINDER¹

RÉSUMÉ

Nous examinons dans cet article divers aspects de la conception et de l'analyse des études dans le cadre desquelles les mêmes unités sont examinées à différents moments, au fil des ans. Ces études incluent les enquêtes longitudinales, ainsi que les analyses longitudinales d'études rétrospectives et de données administratives ou données du recensement. Nous nous intéressons ici tout particulièrement aux problèmes particuliers qui résultent du caractère longitudinal de l'étude, en examinant quatre des principales composantes de l'étude longitudinale, à savoir la conception, la mise en oeuvre, l'évaluation et l'analyse. Une attention particulière doit être portée à chacune de ces étapes durant la planification d'une étude longitudinale. Au nombre des questions qui sont liées à la nature longitudinale des études, mentionnons les suivantes: concepts et définitions, bases de sondage, méthodes d'échantillonnage, collecte des données, traitement de la non-réponse, imputation, validation des données ainsi que analyse et diffusion des données. En présupposant que les exigences fondamentales liées à la conduite d'une enquête transversale sont connues, nous énonçons ci-après les questions et les problèmes qui se dessinent pour bon nombre d'études longitudinales.

MOTS CLÉS: Bases de sondage; données administratives; collecte des données; non-réponse; imputation; estimation; analyse des données.

1. RAISONS JUSTIFIANT LA CONDUITE D'UNE ÉTUDE LONGITUDINALE

Chaque année, des milliers d'enquêtes sont menées par des organismes statistiques à travers le monde. Ces enquêtes visent habituellement à recueillir de l'information pour étayer la prise de décisions ou l'élaboration de politiques. Ces enquêtes ne sont pas menées uniquement à des fins historiques, mais aussi pour recueillir de l'information sur les mesures que l'on pourrait prendre pour faciliter la mise en oeuvre de divers changements de politiques. La plupart des enquêtes sont basées sur des données transversales, c'est-à-dire sur des données obtenues dans le cadre d'une enquête réalisée auprès d'une population donnée, à un moment précis. Des sommaires sont également établis sur la population à l'étude, au moment de l'enquête. Cependant, il arrive très souvent que le but d'une enquête soit, non pas de brosser le tableau de la situation réelle au moment de l'enquête, mais plutôt de connaître les répercussions de divers changements envisagés. Ou encore, il se peut que l'on veuille surveiller les effets d'une modification de politique dont l'entrée en vigueur est prévue prochainement. Le facteur temps est ici l'élément déterminant. Lorsqu'on cherche à obtenir de l'information sur certains phénomènes, par exemple sur l'état de santé ou le niveau de scolarité d'une population, il faut alors examiner les divers déterminants de ces phénomènes. Cependant, il arrive parfois que la relation temporelle elle-même ne soit pas claire, c'est-à-dire quant aux causes ayant précédé les effets. Ces causes pourraient être mesurées si, plutôt que de faire une enquête transversale, les enquêtes

L'avantage de concevoir une étude longitudinale vient de ce que la même méthodologie peut être utilisée pour chacun des cycles de l'enquête, ce qui en retour peut accroître la validité des conclusions qui en seront tirées. Souvent, la conduite d'enquêtes répétées auprès des mêmes répondants s'avère le meilleur moyen de comprendre les profils des changements économiques et sociaux. Il existe une autre approche, moins souhaitable mais néanmoins susceptible de donner des résultats satisfaisants, qui consiste à répéter l'enquête sans nécessairement faire appel aux mêmes répondants, ce qui peut réduire les coûts. L'élément

La profitureation des études longitudinales s'explique également du fait que les sources de données administratives peuvent aujourd'hui être utilisées plus efficacement, ce qui rend faisables certaines études longitudinales. De plus en plus de données administratives sont disponibles. Ces données sont souvent recueillies automatiquement auprès des mêmes personnes, durant une période de temps. Par conséquent, même si les données obtenues de sources administratives ne sont pas idéales, elles peuvent néanmoins constituer une bonne source d'information substitutive pour les enquêtes.

étaient échelonnées sur une certaine période, sous forme d'une série d'enquêtes transversales ou encore en utilisant le même panel de répondants, d'un cycle d'enquête à un autre. C'est cette notion basée sur le bon sens qui est à l'origine de la popularité croissante des études longitudinales. Ces dernières offrent en outre l'avantage de réduire l'importance des effets des variables non observées, lorsque les différences dans le temps sont définies en comparant les données recueillies auprès des mêmes répondants.

1

David A. Binder, Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Immeuble R.H. Coats, Parc Tunney, Ottawa, (Ontario), Canada, K1A 0T6.

Sinclair et Gastwirth étudient le problème des erreurs de classification en ce qui concerne les situations d'activité dans la Current Population Survey de la U.S. Bureau of the Census. À cette fin, ils adaptent la méthode mise au point par Hui et Walter, qui convient bien aux données dichotomiques utilisant des données de réinterview, à des données trichotomiques. Contrairement aux autres méthodes, celle-ci ne présuppose pas que les données de réinterview sont dépourvues d'erreurs, mais part plutôt de l'hypothèse qu'il existe une erreur à la fois dans les données de l'interview initiale et dans celles de la réinterview. Les auteurs font une estimation empirique en comparant les taux d'erreurs estimés générés par leur méthode à ceux des autres méthodes existantes telles que celle de Poterba et Summers, et arrivent à la conclusion que le degré de sous-estimation de l'erreur tend à être plus élevé lorsque le taux véritable de chômage est effectivement élevé. Finalement, plutôt que de supposer un taux d'erreur constant d'un bout à l'autre, ils tentent d'effectuer une analyse qui présume que les taux d'erreur ne sont constants qu'au sein de regroupements temporels associés à divers niveaux de chômage.

Renssen examine la difficulté de combiner des informations sur des variables tirées de deux grandes enquêtes distinctes, au moyen des informations auxiliaires provenant d'une troisième enquête de moindre importance qui recueille toutes les variables. En se fondant sur des concepts comme l'appariement statistique et l'échantillonnage, il propose des méthodes pour la production de tableaux à double entrée, la production de fichiers de microdonnées et l'estimation des corrélations. En ce qui concerne la production des tableaux à double entrée, sa démonstration conduit à l'étude de deux ensembles différents de contraintes liées à l'échantillonnage, l'un appelé stratification incomplète à double entrée et le deuxième, stratification synthétique à double entrée. Lors d'une étude de simulation utilisant les données tirées d'une étude pilote effectuée pour l'Enquête-ménage sur les conditions de vie aux Pays Bas, le calage fondé sur la stratification synthétique à double entrée se révèle bien supérieure.

Arnab évalue différentes stratégies d'échantillonnage en deux occasions. L'échantillon utilisé lors de la deuxième occasion est une combinaison d'un sous-échantillon du premier et d'un nouvel échantillon non apparié. Diverses stratégies de sous-échantillonnage du premier échantillon et d'estimation d'un total lors de la deuxième occasion sont mises en comparaison. Il examine les stratégies existantes décrites dans la documentation et en propose deux nouvelles. L'efficacité des diverses stratégies est comparée à la fois analytiquement et empiriquement.

Finalement, Korn et Graubard examinent le problème de la production d'intervalles de confiance pour les proportions comportant un dénominateur positif prévu faible. Signalant au passage que les intervalles binomiaux de Clopper-Pearson utilisés traditionnellement dans les contextes hors enquête sont inappropriés lorsqu'il est question de les utiliser avec des données d'enquête complexes, ils proposent une modification de ces intervalles. Par des simulations, ils comparent ensuite les intervalles proposés aux autres qui sont communément utilisés, dont les intervalles de transformation logit, les intervalles de Breeze (1990) fondés sur une approximation de Poisson et des intervalles linéaires fondés sur la normalité. Ils illustrent aussi les trois autres méthodes proposées avec des applications utilisant des données tirées de la National Health and Nutrition Examination Survey et la Hispanic Health and Nutrition Examination Survey.

Le rédacteur en chef

Dans ce numéro

Le présent numéro de *Techniques d'enquête* commence par une section spéciale intitulée «Enquêtes longitudinales et analyse» qui contient six des communications présentées lors de la Réunion satellite de l'AISE-AISO sur les études longitudinales qui s'est tenue à Jérusalem en 1997. Une ou deux communications présentées lors de cette conférence, qui n'ont pu être prêtes à temps pour être publiées dans ce numéro, le seront peut-être ultérieurement. Je suis très reconnaissant à Gad Nathan et à Christopher Skinner d'avoir agi à titre de rédacteurs coordonnateurs pour cette section spéciale. Sans leur ténacité et leur ardeur au travail, ce projet n'aurait jamais vu le jour.

La première communication de la section spéciale, signée par Binder, introduit le sujet en passant en revue l'état actuel de la situation et les défis qui se rattachent aux enquêtes longitudinales par comparaison aux enquêtes transversales. La discussion se divise en quatre parties, examinant tout à tour les problèmes particuliers et les défis qui se posent lors de la conception, de la mise en oeuvre, de l'évaluation et de l'analyse des enquêtes longitudinales.

Bassi, Torelli et Trivellato étudient le problème de l'estimation des flux bruts entre les situations d'activité lorsque les données sont entachées d'erreurs de classification. Ils examinent d'abord diverses stratégies pour recueillir des données longitudinales sur la situation par rapport au marché du travail, et les possibilités qu'elles pourraient offrir en ce qui concerne les erreurs de classification. Ensuite, ils présentent un cadre général de modélisation et un modèle «LSRBL modifié» visant à ajuster les estimations de flux bruts afin de corriger les erreurs de classification. Les méthodes sont illustrées au moyen de deux études de cas utilisant des données tirées de la U.S. Survey of Income and Program Participation et de l'Enquête française sur la population active.

Clarke et Chambers étudient l'incidence de la non-réponse dont il faut tenir compte au niveau du ménage. Ils proposent une catégorie de modèles pour la non-réponse dont il faut tenir compte au niveau du ménage. Ils utilisent ensuite des simulations afin de démontrer que les estimations des flux bruts de la population active peuvent être biaisées en présence de cette non-réponse dont il faut tenir compte au niveau du ménage, et que les estimations en fonction de modèles de non-réponse au niveau du ménage peuvent permettre de réduire ce biais. Lorsque le mécanisme de non-réponse au niveau du ménage est correctement spécifié, la source de biais est complètement éliminée; toutefois, même les modèles de non-réponse au niveau du ménage incorrectement spécifiés peuvent réduire ce biais.

Salamin examine le problème de l'estimation d'une variation proportionnelle pour une petite région. Il illustre comment un modèle de régression logistique multidimensionnelle général peut être utilisé pour décrire les données longitudinales obtenues à partir d'un plan d'échantillonnage à renouvellement de panel. Il étudie aussi comment les paramètres de ce modèle peuvent être limités afin de décrire divers types de dépendance entre les observations répétées, ce qui entraîne d'autres estimations des variations fondées sur le modèle. La méthode est illustrée par l'estimation de variations dans la probabilité de trouver du travail dans un canton suisse au moyen de données tirées de l'Enquête suisse sur la population active (ESPA). Par comparaison à de simples différences dans les proportions estimées de personnes occupant un emploi, les estimations fondées sur le modèle comportent de plus petites erreurs types.

Dorfman, dans sa communication, tente de traiter les indices de prix à la consommation d'un point de vue statistique. Il commence par passer en revue la théorie des indices de prix en général, décrivant la méthode stochastique et les objections qui sont formulées à son endroit. Ensuite, il propose une modification de cette méthode stochastique en fonction d'une modélisation à espace d'états qui contourne la majorité des critiques formulées contre celle-ci. La méthode est illustrée au moyen de données sur les prix et les quantités du thon en boîte.

Dans la dernière communication de la section spéciale, Tambay, Schiopu-Kratina, Mayda, Stukel et Nadon décrivent le traitement de la non-réponse dans l'Enquête nationale sur la santé de la population. Les données recueillies au cours du premier cycle de l'enquête sont considérées comme des prédicteurs éventuels de la non-réponse au cours du cycle deux. Un algorithme DAICH (détection automatique d'interactions du chi carré) est utilisé afin de déterminer les classes de pondération pour la correction de la non-réponse au cours du cycle deux. La communication décrit aussi brièvement le plan d'échantillonnage suivi et les autres étapes du calcul des poids d'estimation.

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Volume 24, numéro 2, décembre 1998

TABLE DES MATIÈRES

Dans ce numéro	105
Enquêtes longitudinales et analyse Rédacteurs-coordonateurs Gad Nathan et Christopher Skinner	
D. A. BINDER Les enquêtes longitudinales: Pourquoi ces enquêtes différent-elles de toutes les autres enquêtes?	107
F. BASSI, N. TORELLI et U. TRIVELLATO Stratégies de collecte de données et de modélisation dans l'estimation de flux bruts relatifs à la population active entachés d'erreurs de classification	117
P. S. CLARKE et R. L. CHAMBERS Estimation des flux bruts de la population active provenant d'enquêtes donnant lieu à une non-réponse dont il faut tenir compte au niveau du ménage	133
P.-A. SALAMIN Analyse longitudinale de données de l'enquête suisse sur la population active par régression logistique multidimensionnelle	141
A. H. DORFMAN Les enquêtes sur les indices de prix en tant qu'études quasi-longitudinales	151
J.-L. TAMBAY, I. ŞCHIOPU-KRATINA, J. MAYDA, D. STUKEL et S. NADON Traitement de la non-réponse du cycle deux de l'enquête nationale sur la santé de la population	159
M. D. SINCLAIR et J. L. GASTWIRTH Estimations des erreurs de classification dans l'enquête sur la population active et analyse de leur incidence sur les taux de chômage publiés	171
R. H. RENNSSEN Utilisation de méthodes d'appariement statistique dans l'estimation de calage	185
R. ARNAB Échantillonnage en deux cycles: Estimation de la population totale	201
E. L. KORN et B. I. GRAUBARD Intervalles de confiance pour les proportions à petit nombre d'événements positifs prévus estimées au moyen des données d'enquête	209
Remerciements	219

TECHNIQUES D'ENQUÊTE

Une revue éditée par Statistique Canada

Techniques d'enquête est répertoriée dans The Survey Statistician, Statistical Theory and Methods Abstracts et SRM Database of Social Research Methodology, Erasmus University. On peut en trouver les références dans Current Index to Statistics, et Journal Contents in Qualitative Methods.

COMITÉ DE DIRECTION

Président G.J. Brackstone
Membres D. Binder
G.J.C. Hole
F. Mayda (Directeur de la Production)
C. Patrick

COMITÉ DE RÉDACTION

Rédacteur en chef M.P. Singh, *Statistique Canada*
Rédacteurs associés

D.R. Bellhouse, *University of Western Ontario*
D. Binder, *Statistique Canada*
J.-C. Deville, *INSEE*
J.D. Drew, *Statistique Canada*
J. Eltinge, *Texas A&M University*
W.A. Fuller, *Iowa State University*
R.M. Groves, *University of Maryland*
M.A. Hidiroglou, *Statistique Canada*
D. Holt, *Central Statistical Office, U.K.*
G. Kalton, *Westat, Inc.*
R. Lachapelle, *Statistique Canada*
P. Lahiri, *University of Nebraska-Lincoln*
S. Linacre, *Australian Bureau of Statistics*
G. Nathan, *Central Bureau of Statistics, Israel*

D. Pfeiffermann, *Hebrew University*
J.N.K. Rao, *Carleton University*
L.-P. Rivest, *Université Laval*
I. Sande, *Bell Communications Research, U.S.A.*
F.J. Scheuren, *Ernst and Young, LLP*
J. Sedransk, *Case Western Reserve University*
R. Sitter, *Simon Fraser University*
C.J. Skinner, *University of Southampton*
R. Valliant, *Westat, Inc.*
V.K. Verma, *University of Essex*
P.J. Waite, *U.S. Bureau of the Census*
J. Waksberg, *Westat, Inc.*
K.M. Wolter, *National Opinion Research Center*
A. Zaslavsky, *Harvard University*

POLITIQUE DE RÉDACTION

Techniques d'enquête publie des articles sur les divers aspects des méthodes statistiques qui intéressent un organisme statistique comme, par exemple, les problèmes de conception découlant de contraintes d'ordre pratique, l'utilisation de différentes sources de données et de méthodes de collecte, les erreurs dans les enquêtes, l'évaluation des enquêtes, la recherche sur les méthodes d'enquête, l'analyse des séries chronologiques, la désaisonnalisation, les études démographiques, l'intégration de données statistiques, les méthodes d'estimation et d'analyse de données et le développement de systèmes généralisés. Une importance particulière est accordée à l'élaboration et à l'évaluation de méthodes qui ont été utilisées pour la collecte de données ou appliquées à des données réelles. Tous les articles seront soumis à une critique, mais les auteurs demeurent responsables du contenu de leur texte et les opinions émises dans la revue ne sont pas nécessairement celles du comité de rédaction ni de Statistique Canada.

Présentation de textes pour la revue

Techniques d'enquête est publiée deux fois l'an. Les auteurs désirant faire paraître un article sont invités à faire parvenir le texte rédigé en anglais ou en français au rédacteur en chef, M. M.P. Singh, Division des méthodes d'enquêtes auprès des ménages, Statistique Canada, Tunney's Pasture, Ottawa (Ontario), Canada K1A 0T6. Prière d'envoyer quatre exemplaires dactylographiés selon les directives présentées dans la revue. Ces exemplaires ne seront pas retournés à l'auteur.

Abonnement

Le prix de Techniques d'enquête (n° 12-001-XPB au catalogue) est de 47 \$ par année au Canada et de 47 \$ US par année à l'extérieur du Canada. Prière de faire parvenir votre demande d'abonnement à Statistique Canada, Division des opérations et de l'intégration, Gestion de la circulation, 120, avenue Parkdale, Ottawa (Ontario), Canada K1A 0T6 ou commandez par téléphone au (613) 951-7277 ou au 1 800 700-1033, par télécopieur au (613) 951-1584 ou au 1 800 889-9734 ou par Internet : order@statcan.ca. Un prix réduit est offert aux membres de l'American Statistical Association, l'Association Internationale de Statisticiens d'Enquête, l'American Association for Public Opinion Research et la Société Statistique du Canada.

Ottawa

ISSN 0714-0045

Périodicité: semestrielle

N° 12-001-XPB au catalogue

Janvier 1999

Tous droits réservés. Il est interdit de reproduire ou de transmettre le contenu de la présente publication, sous quelque forme ou par quelque moyen que ce soit, enregistré ou non, sur support magnétique, reproduction électronique, mécanique, photographique, ou autre, ou de l'emmagasiner dans un système de recouvrement, sans l'autorisation écrite préalable des Services de concession des droits de licence, Division du marketing, Statistique Canada, Ottawa, Ontario, Canada K1A 0T6.

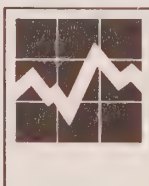
© Ministère de l'Industrie, 1999

Publication autorisée par le ministre
responsable de Statistique Canada

DÉCEMBRE 1998 • VOLUME 24 • NUMÉRO 2

UNE REVUE ÉDITÉE PAR STATISTIQUE CANADA

TECHNIQUES D'ENQUÊTE



4427



NUMÉRO 2

•

VOLUME 24

•

DÉCEMBRE 1998

UNE REVUE
ÉDITÉE
PAR STATISTIQUE CANADA

N° 12-001-XPB au catalogue

TECHNIQUES D'ENQUÊTE



